TU-Net: Transformer based U-Net for left ventricle MRI segmentation

Amit Pandey[1], Akansha Singh[1], Ajith Abraham[2] and Krishna Kant Singh[3]

[1] School of CSET, Bennett University, Gautam Buddha Nagar, India
[2] School of AI, Bennett University, Gautam Buddha Nagar, India
[3] Delhi Technical Campus, Greater Noida, India

Abstract

Accurate segmentation of the left ventricle in cardiac MRI images is crucial for evaluating cardiac function and diagnosing cardiovascular conditions. Traditional approaches, including the commonly used U-Net architecture, struggle with capturing the global contextual information required for precise segmentation. This study introduces U-Net MHSA, an enhanced version of U-Net that incorporates Multi-Head Self-Attention (MHSA) in the bottleneck layer to overcome these limitations. By combining the strengths of convolution layers and attention mechanisms, our model effectively captures long-range dependencies while preserving spatial coherence. Our model U-Net MHSA gives better results as compared to the baseline U-Net on the MICCAI 2009 Left Ventricle Segmentation Challenge dataset. U-Net MHSA gives higher scores as compared to baseline U-Net in terms of precision 0.799531 and accuracy 0.797943. Although the model gives a minor trade-off with slightly reduced recall and Intersection over Union (IoU). The overall results shows that the integration of MHSA with U-Net architecture improves the medical image segmentation.

Keywords

MRI, Cardiac function, U-Net, Multi-Head Self-Attention, medical image segmentation, Self-Attention

1. Introduction

Medical image segmentation (MIA) [1] plays a crucial role in modern healthcare, where accurate and precise diagnostic tools for example Magnetic Resonance Imaging (MRI), X-ray, and CT scans [2] are very crucial in clinical decision-making. Traditional methods like manual and semi-automatic segmentation are purely based on human inputs and are not so much accurate and precise but also time-consuming. In the last few years machine learning [3], deep learning [4], and convolutional neural network [5] have revolutionized the medical image field. U-Net [6], based on a convolutional neural network came into the picture in 2015 and revolutionized the field of medical imaging due to its unique U-Shaped architecture and skip connections. By using skip connections U-Net concatenates the low-level features with high-level features for more accurate and precise segmentations of medical images. Despite having a lot of advantages and success U-Net has some limitations also. Initial layers of the encoder path have poor representations of feature maps and these feature maps also pass through skip connections, which have no use and also increase the time and space complexity. U-Net was also not able to handle long-rage dependencies and parallel computations. In order to handle these limitations, we propose TU-Net a hybrid model which integrates MHSA [7] with U-Net architecture in bottleneck. TU-Net aims to use the strengths of both architectures and gives better performance by capturing global image context and also retains fine-grained spatial feature, which is essential for accurate and precise segmentation. In further sections we explain in detail self-attention, MHSA block and U-Net architecture.

ProfIT AI 2024: 4th International Workshop of IT-professionals on Artificial Intelligence (ProfIT AI 2024), September 25–27, 2024, Cambridge, MA, USA

: e21soep0035@bennett.edu.in (A. Pandey); akansha1.singh@bennett.edu.in (A. Singh); ajith.abraham@bennett.edu.in (A. Abraham); Krishnaiitr2011@gmail.com (K.K. Singh)

0009-0000-1317-952X (A. Pandey); 0000-0002-5520-8066 (A. Singh); 0000-0002-0169-6738 (A. Abraham); 0000-0002-6510-6768 (K.K. Singh)

2. Methodology

In this particular section, we explain the methodology used in developing TU-Net, a novel architecture that improves the performance of the U-Net baseline model with Transformer-based Multi-Head Self-Attention (MHSA) for left ventricle MRI segmentation. The steps of our model are shown in Figure 1. In the first step, the input image passes into the encoders after that in the second step output of the last encoder passes into the MHSA block and finally output of the MHSA block passes into decoders and gets the output segmentation map.

2.1. U-Net Architecture with Integration of MHSA Block

The U-Net architecture as shown in Figure 2 was famous for its unique U-shaped encoder-decoder architecture, enabling precise localization and segmentation capabilities. In the encoder path, feature maps are extracted by two successive 3x3 convolutions followed by ReLU activation functions. After that 2x2 max-pooling operations are used to down-sample the image size. The above process is repeated five times as five encoders are used in U-Net architecture. After the fifth encoder in the bottleneck section, we integrate the MHSA module which processes the feature maps received from the last encoder and enables the proposed architecture to capture global contexts and long-range dependencies within the image. Conversely, in the decoder path, feature maps are up-sampled by using 2x2 convolutions, and after that concatenate the feature maps from the corresponding encoder side with the decoder side. After this step two successive 3x3 convolution operations were used followed by the ReLU activation function this process was also repeated five times and finally 1x1 convolution operation was used after the last decoder to give the final segmentation map.

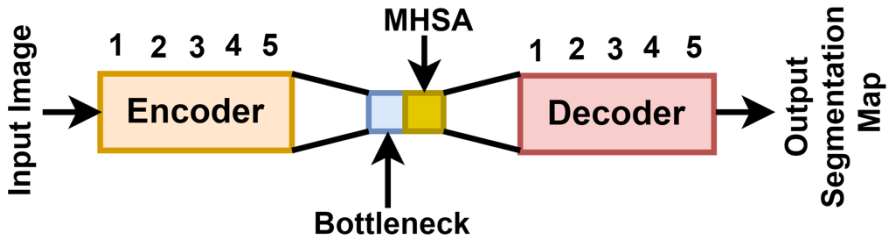

Figure 1: Steps of Proposed Model

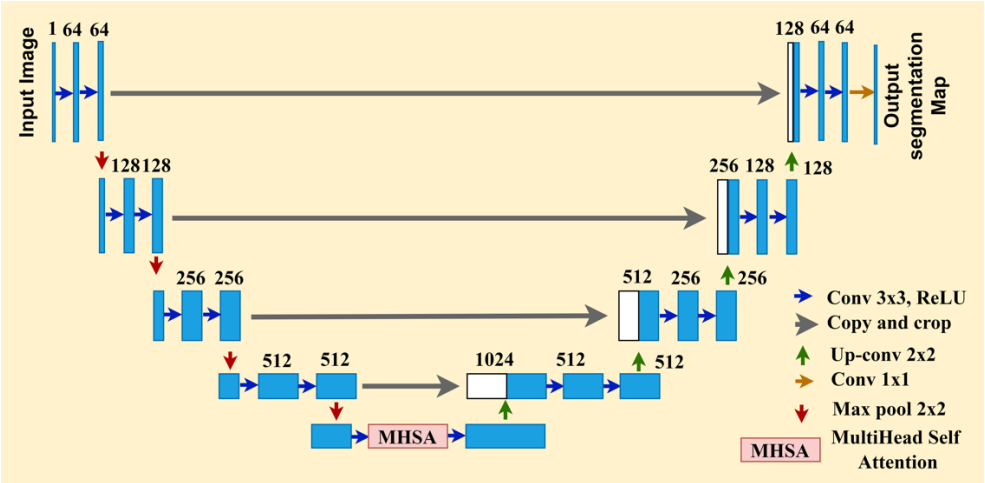

Figure 2: U-Net with MHSA

2.2. Multi-Head Self-Attention (MHSA)

MHSA is an advanced technique used in transformer models to improve their ability to process information. Instead of relying on a single attention mechanism with queries, keys, and values all having dimensionality u_{model}, MHSA divides this process into multiple, parallel attention operations. Each of these operations, known as heads, maps the queries, keys, and values into smaller dimensions u_k and u_v using distinct learned linear projections. Attention is computed in parallel for each head, and the resulting outputs, which are u_v-dimensional, are concatenated and re-projected to produce the final output. This approach allows the model to focus on various representation subspaces at different positions, whereas a single attention head would average these aspects together.

Overcome U-Net's limitation in capturing long-range dependencies, we incorporated MHSA into the bottleneck of the U-Net architecture. MHSA, which a concept derived from transformers, allows the model to attend to various parts of the input image simultaneously, thereby capturing global context more effectively as mentioned in Figure 1. The TU-Net architecture retains the basic structure of U-Net but integrates MHSA in the bottleneck layer to enhance its ability to capture global information. The self-attention mechanism as shown in Figure 4 works by calculating attention scores between various positions within the input image. It consists of three main components: Query (Q), Key (K), and Value (V). The attention scores A are calculated by taking the scaled dot-product of Q and K, and then applying a Soft-Max function to obtain the attention weights, as shown in equation 1.

$$A = softmax(\frac{QK^T}{\sqrt{d_k}}) \tag{1}$$

where d_k is the dimensionality of the key vectors. These weights are then applied to V for the final output, as shown in Equation 2.

$$\text{Attention}(Q, K, V) = A \cdot V \tag{2}$$

This process is performed multiple times in parallel to create MHSA, enabling the model to simultaneously focus on different regions of the image as illustrated in Figure 3. The step-wise working of MHSA is shown in Figure 5. From left to right. In the first step, we simply pass the input sequence, In the second step, we embed each word, In all encoders except encoder 0, we don't need embedding. In the third step, we split into eight heads and multiplied X or R with weight matrices. In the fourth step, calculate the attention scores by making use of Q, K, and V matrices. In the final step, concatenate the results of Z matrices and then multiply with the weight matrix W^0 and finally produce the output.

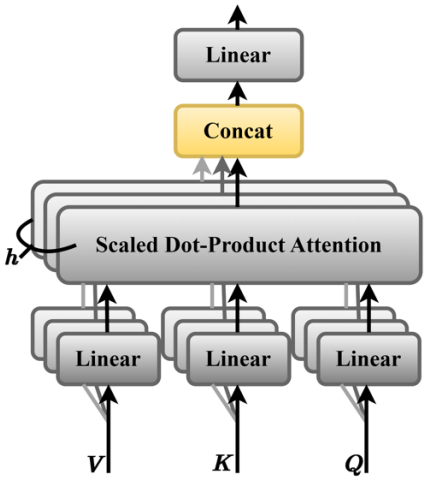

Figure 3: Multi-head self-attention (MHSA)

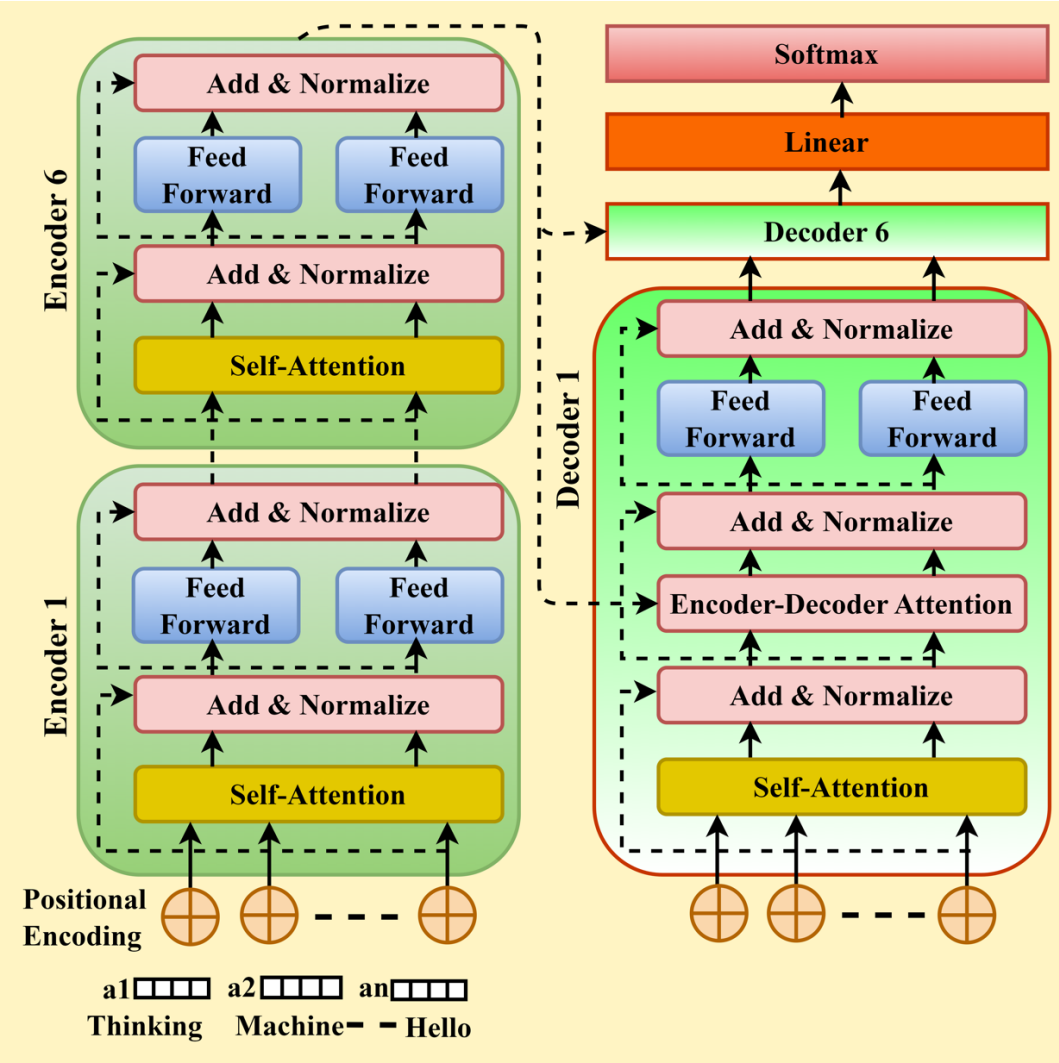

Figure 4: Detailed Architecture of Transformer

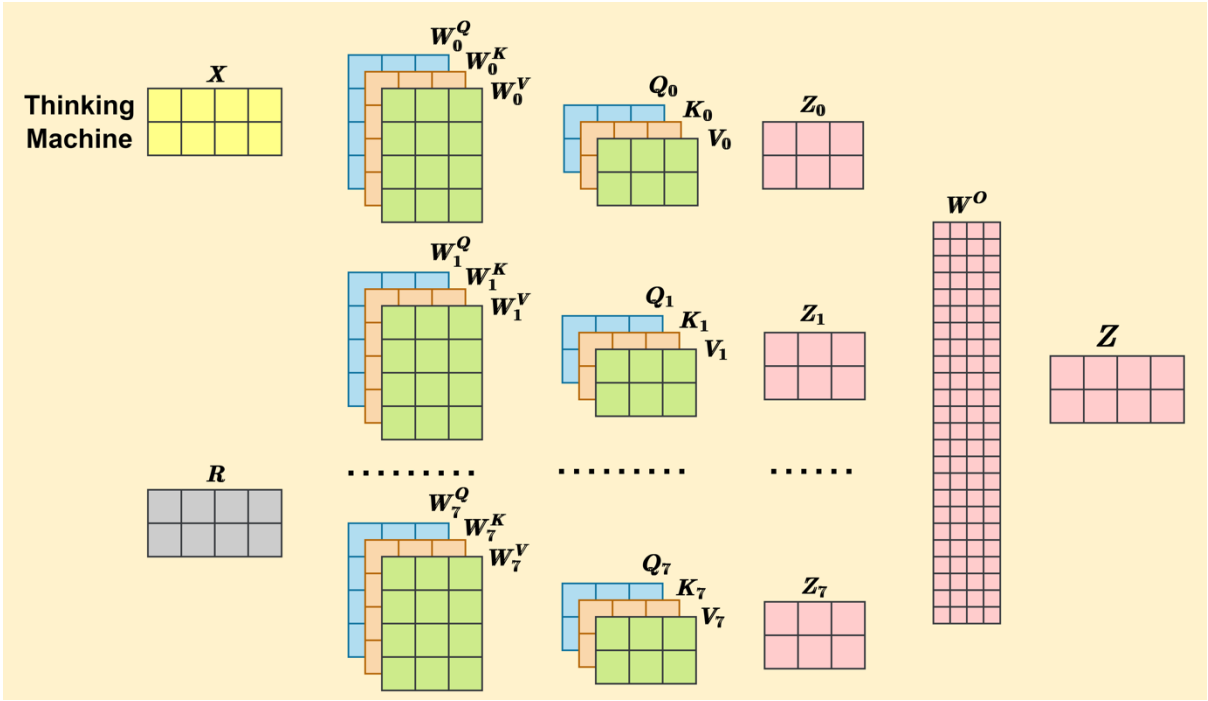

Figure 5: Detailed Working process of MHSA Module

Table 1
Hyperparameters

Hyperparameter	Value
Image Size	256 x 256
Batch Size	64
Epochs	50
Training Images	4900
Validation Images	500
Test Images	266
Total Parameters	48,195,073 (183.85 MB)
Trainable Parameters	48,183,297 (183.80 MB)
Non-trainable Parameters	11,776 (46.00 KB)

2.3. Training Procedure

The training procedure for the TU-Net architecture uses the MICCAI 2009 [8] Left Ventricle Segmentation Challenge dataset. The details of the data set are mentioned in TABLE 1. Before training, the MRI images were subjected to several preprocessing steps to ensure uniformity and enhance model performance. Each image was resized to 256 x 256 pixels, and also the pixel intensity values were normalized. To prevent overfitting, data augmentation techniques such as random rotations, shifts, flips, and zooms were applied to the training dataset. Adam optimizer were used to train the TU-Net model, which is known for its efficiency and capability to handle sparse gradients. A hybrid loss function, which combines binary cross-entropy and Dice was employed to balance pixel-wise accuracy with the overlap between ground truth and predicted masks. During training, the TU-Net model's parameters were iteratively adjusted to minimize the loss function through forward and backward propagation steps. In the forward pass, the input images were fed through the model to obtain predictions, which were then compared to the ground truth masks to compute the loss. In the backward pass, the computed loss was used to update the model parameters through the Adam optimizer.

The model's performance was validated on the 500-image validation set after each epoch, providing insights into its generalization capability on unseen data. This validation process also guided the tuning of hyperparameters. After training concluded, the final model underwent evaluation using a test set comprising 266 images to gauge its performance in real-world scenarios. The final TU-Net model, incorporating MHSA in the bottleneck layer, comprised a total of 48,195,073 parameters, with 48,183,297 being trainable and 11,776 non-trainable, resulting in a model size of 183.85 MB. The training procedure ensured that the model was well-optimized for accurate and reliable segmentation of the left ventricle in MRI images as mentioned in Table 1.

2.4. Evaluation Metrics

The performance evaluation encompasses key metrics including precision, recall, specificity, intersection over union (IoU), and a custom evaluation metric derived from the evaluate generator function, offering a comprehensive assessment of overall accuracy.

3. Results

The performance of the TU-Net model with Multi-Head Self-Attention (MHSA) was evaluated against the standard U-Net model using several key metrics: Precision, Recall, Specificity, IoU, and Accuracy. The evaluation was conducted on the MICCAI 2009 Left Ventricle Segmentation Challenge dataset, focusing on the segmentation of the left ventricle in MRI images. Table 2 below summarizes the comparative results of the two models.

Table 2
Performance Comparison

Model	Precision	Recall	Specificity	IoU	Accuracy
U-Net	0.773880	0.653408	0.996921	0.548658	0.710639
U-Net MHSA	0.799531	0.576392	0.997670	0.503610	0.797943

Precision was higher for the U-Net MHSA model (0.799531) compared to the standard U-Net model (0.773880). This indicates that the incorporation of MHSA helped in reducing false positives. Recall was higher for the U-Net model (0.653408) compared to the U-Net MHSA model (0.576392). This suggests that while the U-Net MHSA model had fewer false positives, it also had a slightly higher number of false negatives. Specificity was slightly better for the U-Net.

MHSA model (0.997670) compared to the standard U-Net model (0.996921). This improvement, albeit small, indicates a better performance in correctly identifying negative samples. The IoU metric was slightly lower for the U-Net MHSA model (0.503610) compared to the standard U-Net model (0.548658). This suggests that the standard U-Net had a slightly better spatial overlap between the predicted and true segmentation masks. Accuracy, evaluated using the evaluate generator function, was significantly higher for the U-Net MHSA model (0.797943) compared to the standard U-Net model (0.710639). This indicates that the overall performance and correctness of the U-Net MHSA model in segmenting the left ventricle were superior.

In addition to the tabular results, Figure 6 illustrates a comparative graph which visually represents the performance disparities between the convolutional U-Net model and the U-Net MHSA model. This graph highlights the enhanced accuracy and precision of the U-Net MHSA model, despite a trade-off in recall and IoU. Figure 7 illustrates a visual comparison between U-Net MHSA and U-Net.

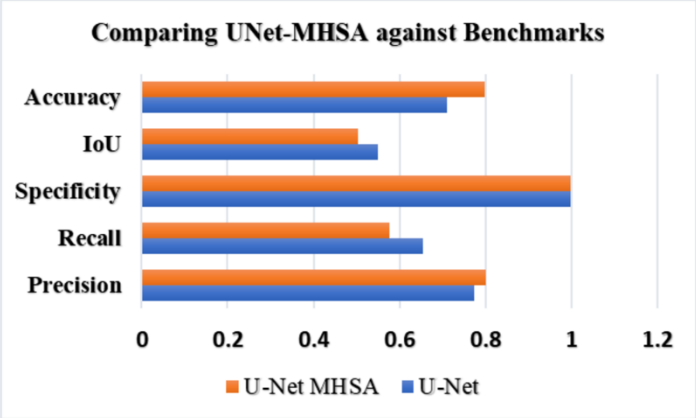

Figure 6: Detailed Working process of MHSA Module

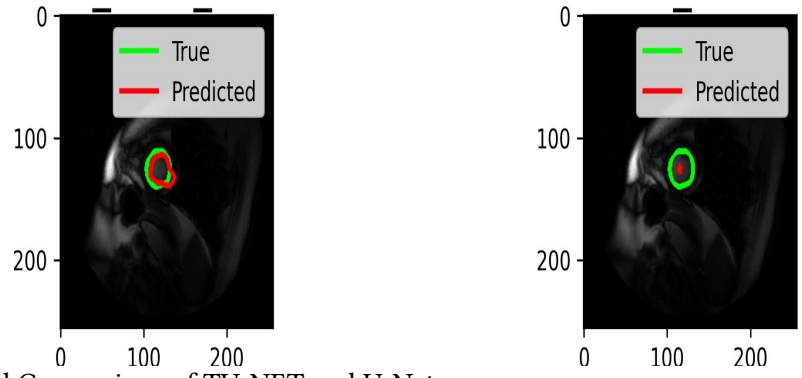

Figure 7: Visual Comparison of TU-NET and U-Net

4. Discussion

This study aimed to enhance the U-Net architecture for medical image segmentation by incorporating MHSA into its bottleneck layer. The results indicate that the enhanced model, U-Net MHSA, shows considerable improvements compared to the standard U-Net, especially regarding precision and overall accuracy. Integrating MHSA into the U-Net framework enables the model to more effectively capture long-range dependencies and contextual relationships within the image, which are crucial for precise segmentation. Our findings show that U-Net MHSA achieved a precision of 0.799531 and an accuracy of 0.797943, outperforming the standard U-Net, which had a precision of 0.773880 and an accuracy of 0.710639. These enhancements highlight the benefits of incorporating attention mechanisms to improve the TU-Net's ability to focus on important features throughout the entire image.

However, while U-Net MHSA showed notable gains in precision and accuracy, it did exhibit a slightly lower recall (0.576392) and IoU (0.503610) compared to the standard U-Net, which had a recall of 0.653408 and an IoU of 0.548658. This suggests that although U-Net MHSA is more precise in identifying the left ventricle, it may miss some true positives, leading to a lower recall. The decreased IoU indicates a reduced overlap between predicted and actual segmentations, pointing to a potential area for further optimization. The trade-off between precision and recall observed in our study is a common challenge in segmentation tasks. Precision measures how many of the identified segments are correct, while recall measures how many of the actual segments were identified. Achieving a balance between these metrics is crucial for practical applications, especially in medical imaging, where both false positives and false negatives can have significant consequences. One of the strengths of our approach is the ability of MHSA to capture global context, which is often overlooked by traditional convolution operations that primarily focus on local features. By attending to different parts of the image simultaneously, MHSA provides a more comprehensive understanding of spatial relationships, enhancing the model's ability to delineate complex anatomical structures. The overall higher accuracy of U-Net MHSA highlights its robustness and effectiveness for the task of left ventricle segmentation. The additional computational cost introduced by the MHSA module is justified by the performance gains, demonstrating the potential of self-attention mechanisms in improving convolution neural network architectures.

5. Conclusions

We present U-Net MHSA for medical image segmentation, especially left ventricle in heart images. U-Net MHSA is an advanced architecture, incorporating MHSA into the bottleneck layer has shown significant improvements in precision and overall accuracy. U-Net MHSA has outperformed standard U-Net. While previously standard U-Net had a precision value of 0.733880 and accuracy value of 0.710639, now after integration of U-Net MHSA, the precision value has become 0.799531 and accuracy value has become 0.797943 which is better than before. Along with all these benefits, there is some decrease in recall and Intersection over Union (IOU) values with U-Net MHSA. U-Net MHSA demonstrates the potential of convolution neural network architecture, self-attention mechanism to improve segmentation performance. Future research should focus on optimizing the attention mechanism and validating the model on different segmentation tasks and datasets to ensure its generalizability and robustness in various clinical scenarios.

Acknowledgements

We would like to express our sincere gratitude to the Department of Computer Science at Bennett University for providing the necessary resources and support throughout this research. Special thanks to our colleagues and mentors, whose insights and expertise were invaluable in the development and refinement of this study. This work was not funded. We also extend our appreciation to the MICCAI 2009 Left Ventricle Segmentation Challenge for providing the dataset.

References

[1] Singh, Krishna Kant, and Akansha Singh. "A study of image segmentation algorithms for different types of images." International Journal of Computer Science Issues (IJCSI) 7.5 (2010): 414.

[2] Acharya, Raj, et al. "Biomedical imaging modalities: a tutorial." Computerized Medical Imaging and Graphics 19.1 (1995): 3-25.

[3] Singh, Pushpa, et al. "Diagnosing of disease using machine learning." Machine learning and the internet of medical things in healthcare. Academic Press, 2021. 89-111.

[4] Sharma, Poonam, and Akansha Singh. "Era of deep neural networks: A review." 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2017.

[5] O'Shea, K. "An introduction to convolutional neural networks." arXiv preprint arXiv:1511.08458 (2015).

[6] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. Springer International Publishing, 2015.

[7] Vaswani, A. "Attention is all you need." Advances in Neural Information Processing Systems (2017).

[8] Cardiac MR Left Ventricle Segmentation Challenge. URLhttp://hdl.handle.net/10380/307