Leveraging Gram Matrix Representation in Shallow DNN for GI Tract MRI Image Segmentation

Fariza Siddiqua, Tamim Ahmed and Mohammed Imamul Hassan Bhuiyan
Department of Electrical Electronic Engineering
Bangladesh University of Engineering and Technology
Fariza.eee17.buet@gmail.com, tamim.ahmed.tamim@gmail.com, imamul@eee.buet.ac.bd
Dhaka – 1211, Bangladesh

***Abstract*— GI tract MRI image segmentation is important for the diagnosis and treatment of many diseases. In case of GI tract cancer or tumor treatment, radio oncologists must apply X-ray beams pointing towards the tumor cell while avoiding the other organs. The traditional segmentation process for the MRI scan is time consuming and labor intensive. A computer aided fast and accurate method is required. UNet, UNet++, Attention UNet are some of the deep learning architectures used in this purpose having remarkable performance. Most of these architectures inherit significant computational complexity. In this paper, a gram matrix oriented shallow DNN in proposed which will bypass the traditional attention mechanism. Due to the simplified computation of the proposed model, it yields similar or superior segmentation performance as compared to existing models and it has much lower number of trainable parameters in the range of 10 to 75 percent less the other models.**

Keywords—U-Net, Gram Matrix, GI tract, MRI, Multi-level semantic segmentation

I. INTRODUCTION

Deep learning algorithms have brought about revolutionary change in medical image segmentation. For the treatment and diagnosis of many diseases and abnormalities, segmentation of different organs or parts is a fundamental task. In this paper, we focus on GI tract image segmentation for better and faster treatment. When radio oncologists need to apply X-ray beams to the tumor cells in GI tract, they must segment stomach and intestines from the tumor cells. Manual segmentation in MRI image of GI tract takes about 15 minutes to hours which prolongs the treatment. Therefore, there is high need of automated image segmentation method to save time and make the segmentation more accurate. In case of GI tract image segmentation, this is a type of multi-level semantic segmentation. Deep learning architectures take the lead of this task. UNet [1], Attention UNet [2], Transformer based UNet [3], UNet++ [4], LeViT-UNet++[5], UNet with ResNet encoder [6] are some of the proposed models which can give great segmentation result. However, there are some limitations in these models- (i) UNet [1] performs great in medical image segmentation, but at the cost of large filter size. It requires a large number of trainable parameters. (ii) Attention UNet [2] also performs well but it inherits quadratic computational complexity. (iii) Transformer does patchification of images before segmentation. It causes degradation in segmented images' resolution. (iv) UNet++ [4] has too many layers and skip connections. Its bulk architecture requires expensive hardware to be embedded in. In this paper, we have proposed a Gram matrix-oriented bypass attention mechanism in the U-shaped architecture. It uses less multiplication operation and more concatenation operation. Therefore, the computational complexity becomes almost linear. Because of the implementation of Gram matrix [7], a more developed feature matrix is produced in each step, which makes it possible to reduce the number of layers and the size of filters. Therefore, the trainable parameters are drastically reduced and computation is simplified. For training and validation, we have used MRI scanned GI tract image dataset provided by UW-Madison [8]. It is a three-level semantic segmentation problem where stomach, large bowel, and small bowel from the MRI image. In the result analysis part of this paper, we have compared the performance of our method with UNet [1], UNet++ [4] and ResNet Encoder based UNet[6], LeViT-UNet [5]. The comparison reveals that the proposed method gives a better segmentation performance compared to UNet++ [4] and LeViT UNet++ [5] while requiring significantly reduced number of trainable parameters.

II. RELATED WORKS

A. Medical image Segmentation Based on CNN

Encoder - decoder based U shaped CNN architectures play a pivotal role in medical image segmentation. There are many variants of U-shaped architectures. UNet [1] is one of them which shows remarkable segmentation performance. Among the encoders of UNet, ResNet performed best in case of GI tract MRI image segmentation [6]. To segment small region, weighted ResUNet [9] is used. UNet++ [4] is an elaborate version of UNet which is has less semantic gap. Between the feature maps of encoder and decoder. For 3-D image segmentation, 3-D UNet [10], 3-D Multi ResUNet [11] are used. A modified version of 3-D Multi ResUNet outperformed the traditional 3-D UNet and 3-D Multi ResUNet in case of lung tumor image segmentation [12].

B. *Medical Image Segmentation Based on Transformers*

CNN based models cannot model long range dependencies. Vision Transformer or ViT [13] is used in this task. There are many hybrid networks used in medical image segmentation. These are basically combination of transformer and U shaped architecture. TransUNet [14] is the first hybrid structure. TransUNet [13] is a pioneering framework for medical image segmentation that introduces self-attention mechanisms in a sequence-to-sequence prediction setup. To address the reduction in feature resolution caused by Transformers, it combines a hybrid CNN-Transformer architecture. IT-UNet [15] is used for organ at risk segmentation. For thoraic segmentation, FcTC- UNet [16] is used. Another example of hybrid network in LeViT UNet++ [5] used for GI tract MRI image segmentation. In this method, the usage of vision transformer is implied by combining hard inductive bias properties of CNNs. There are also some purely transformer based methods. Swin Transfomer [17] is one of them. In some case these models performs more accurate.

III. Dataset and Preprocessing

The proposed model is trained and validated with GI tract MRI scanned image dataset provided by UW Madison [8]. The dataset was provided for a competition in Kaggle and only training dataset was available. In the train folder there were subfolders for each patient. The scanned images of each patient were day-wise organized. Each scanned image was in PNG format. For each image 3-levels of segmentation masks were provided and the levels were stomach, large bowel, and small bowel. The masks were in RLE (Run Length Encoded) format. As each image has 3 levels (stomach, large bowel, and small bowel). A percentage distribution of images among stomach, large bowels and small bowels is shown in Fig.1

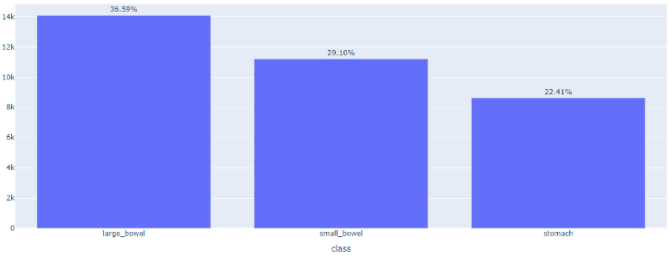

Figure 1: Percentage (%) distribution of training images with mask.

Before training the model with the dataset, the RLE masks were decoded. 70% of the training data was used for training and 30% was used for validation.

IV. Proposed Model

In our proposed model, as shown in Fig.2, we have redesigned the encoder and decoder part of the U-shape architecture. Each layer of the encoder is named as GCFM (Gram matrix concatenated feature matrix) block. There is total 5 GCFM blocks. Each layer in the decoder is named as Expansion block. There is total 5 Expansion blocks. In each GCFM block, a 3-D feature matrix is passed through 4 parallel convolution layers with different kernel size. In each of these convolutional layers, a 1-D vector corresponding to the feature matrix is produced. From four parallel convolutional layers, we get four 1-D vectors. Later these 1-D vectors are concatenated to form the one vector V corresponding to the primary feature matrix. This vector is then transposed. Gram matrix is defined as,

$$\text{Gram matrix} = V \times V^t \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Gram matrix represents the correlation coefficient among the components of the vector V.

This 2-D Gram Matrix is then concatenated in the front and end of the primary 3-D feature matrix. If the size of the feature matrix is 80x80x32, the size of the Gram matrix concatenated feature matrix will be 80x80x34. Gram Matrix represents the correlation of one element of V_1 with the other elements of the same matrix. Therefore, it gives an insight of importance of different parts of the feature matrix. The basic block diagram of the proposed model is given below. Here, five GCFM blocks are used in the encoder side and five Expansion blocks are used in the decoder side.

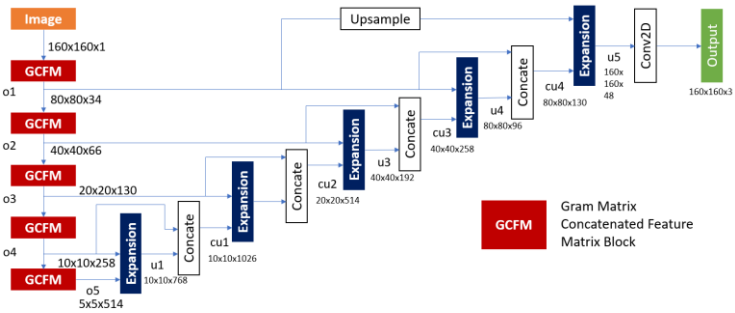

Figure 2: Complete Diagram of the Architecture

A. *GCFM Block:*

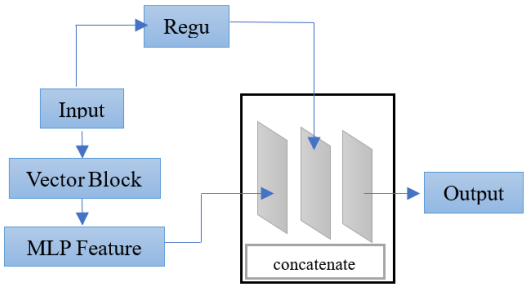

Figure 3: GCFM Block Diagram

In Fig.3 we can see that, GCFM block takes a 3-D tensor as input and sends it to the vector block. The vector block produces the 1-D component vector. MLP feature block takes this vector as input and generates Gram matrix. After that, the Gram Matrix is concatenated with the feature matrix

B. Vector Block:

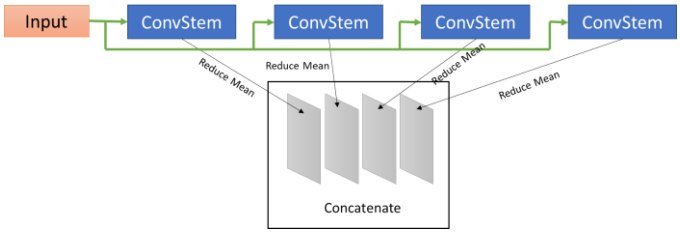

Figure 4: Block Diagram of Vector Block

As seen in Fig.4 the vector block takes the 3-D tensor and sends it to four parallel convstem layers. In each convstem layer, convolution, gelu activation and normalization take place. The output of the convstem layer is channelwise averaged. After passing through the convstem layer, a 1-D vector is generated. From 4 convstem layers, 4 1-D vectors are concatenated and the final 1D component vector is generated.

C. MLP Feature Block:

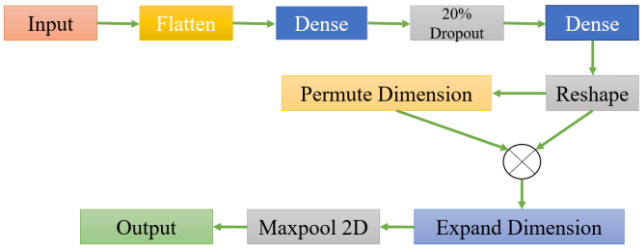

Figure 5: MLP Feature Block

In Fig.5 one can see that, MLP feature block takes the 1-D component vector as input. Then this vector is flattened. After that it is passed through two successive dense convolution layers. The inner product of the vector is determined and the Gram matrix is generated.

D. Expansion Block:

In Fig.6 the expansion block takes two inputs. Considering the first expansion block, the first input is the output of the last GCFM block. And the 2nd input is the output of the 2nd last GCFM block. These two inputs and concatenated after passing through the convolution and transpose convolution layer. 30% dropout is used to avoid overfitting. Then the generated output is again concatenated with the 2nd input. This is operation in one expansion block. This repeats 5 times in out model.

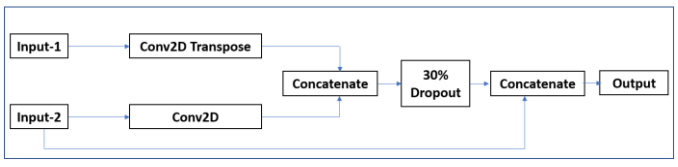

Figure 6: Block Diagram of Expansion Block

V. Result & Analysis

We use around 40 epochs to train the model and for the best use of GPU memory. 70% of the data were used to train the model and the rest of 30% were used to evaluate the model.
For the result analysis, the performance parameter we used are:

$$\text{Dice coefficient} = \frac{2*|\text{A intersect B}|}{(|A|+|B|)} \quad \text{.....(2)}$$

$$\text{IoU(A, B)} = \frac{A \cap B}{|A| + |B|} \quad \text{.....(3)}$$

Dice Loss = 1 - Dice Coefficient.....(4)

BCE_loss = BCE + Dice Loss .....(5)

$$\text{Tversky\_loss}_{\alpha,\beta}(y,\hat{y}) = \frac{TP}{TP+\alpha FP+\beta FN} \quad \text{.....(6)}$$

where, BCE means Binay Cross Entropy and TP, FP and FN mean true positive, false positive and false negative respectively. In (6), α and β are constants. The value of each of these constants are 0.5.

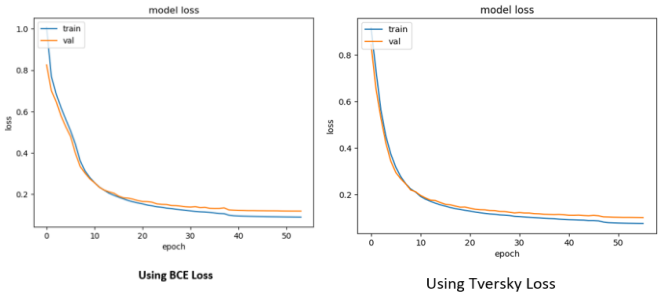

Figure 7: Comparison between BCE loss and Tversky loss

Fig.7 shows the plot of BCE loss and Tversky loss with respect to number of epochs. It is observed that within 40 epochs both the losses decrease significantly.

Table 1: Loss Comparison

Loss	Train	Validation
BCE Loss	0.0975	0.1237
Tversky Loss	0.0765	0.1013

Table.1 gives the minimum losses achieved training and validation data. It is seen that Tversky loss in lower than the BCE loss.

Figure 8: Plot of Dice Coefficient

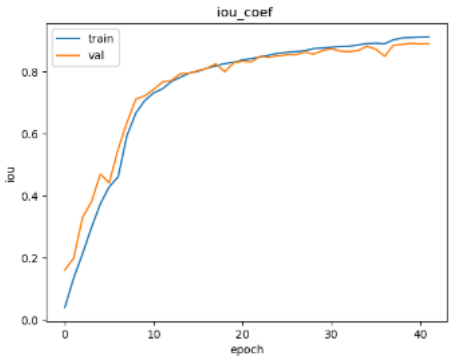

Figure 9: Plot of IoU Coefficient

Likewise, in Fig.8 and Fig.9 the maximum values of Dice coefficient and IoU are obtained within a mere 40 epochs.

Table 2: Comparison with different models

Validation accuracy				
Models	IoU	Dice coefficient	Loss	Parameter Count
Proposed Model	0.912	0.9008	0.1073	9,314,745
UNet	0.9107	0.9032	0.1032	36,893,154
UNet++	0.8764	0.9012	0.1062	10,365,568
LeViT-UNet++ [5]	0.7282	0.7954	0.1347	----------
ResNet Encoded UNet [6]	0.884	----------	---------	----------

Table.2 provides the comparison of the performance of the proposed model with that of generic UNet, generic UNet++, LeViT UNet++ [5] and ResNet encoded UNet [6]. It should be mentioned that for LeViT UNet++ [5], maximum score values are shown among those obtained for four different folds. It can be seen that the IoU of proposed model is 0.14% greater than UNet, 4.06% greater than UNet++, 25.24% greater than LeViT UNet++ [5] and 2.8% greater than ResNet encoded UNet [6]. The dice score of the proposed method is almost same as that of UNet and UNet++ but 13.15% greater than LeViT-UNet++ [5]. The validation loss is also very close to that of UNet and UNet++ and 20.34% less than that of LeViT-UNet++ [5]. If we look at the parameter count, a noticeable improvement can be seen. Our model can perform almost like UNet having 75% less trainable parameters than UNet and 10% less trainable parameters than UNet++. It indicates that our model is much shallower than the other two without doing any sacrifice in case of segmentation performance.

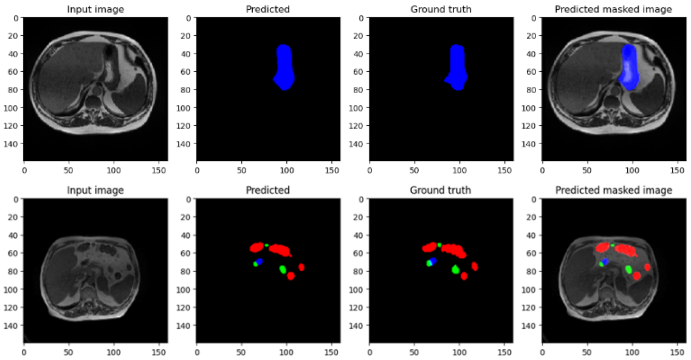

Figure 10: Visual representation of predicted mask and comparison with ground truth; red denotes large bowels, green denotes small bowels and blue denotes stomach

Fig.10 shows the predicted masks for two randomly selected GI tract images. In both cases, the predicted masks are quite close to the ground truths.

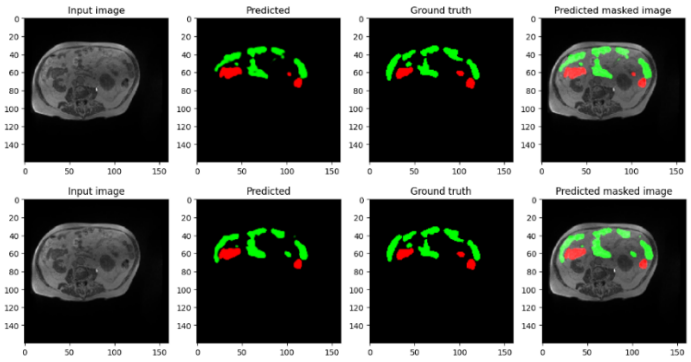

Figure 11: Segmentation comparison between proposed model and generic U-Net (Red denotes large bowel, green denotes small bowel and blue denotes stomach)

Fig.11 shows the comparison of predicted masks by our proposed model and generic U-Net. The first image represents the predicted mask by the proposed model. The 2nd image represents the predicted mask by generic U-Net. It can be seen that, in some portion, U-Net predicted better than our model, while in some portions, our model predicted better than U-Net. Therefore, it can be said that the overall segmentation performance by the proposed method is equivalent to that of U-Net.

VI. CONCLUSION

Segmentation of MRI images of GI tract is an important issue regarding the treatment and prognosis GI tract cancer patients. This paper presents modified UNet architecture incorporating the notion od Gram matrix. It has been shown that, using a publicly available dataset, the proposed architecture can deliver a superior segmentation of stomach, large bowel and small bowel as compared to existing models. Moreover, the proposed model is rather shallow considering significantly reduced number of trainable parameters. This shallow deep neural network has a potential to be embedded in less expensive medical devices and treatment of GI tract cancer can be more feasible.

VII. REFERENCE

[1] Ronneberger, Olaf & Fischer, Philipp & Brox, Thomas. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. LNCS. 9351. 234-241. 10.1007/978-3-319-24574-4_28.

[2] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.

[3] Petit, Olivier & Thome, Nicolas & Rambour, Clément & Themyr, Loic & Collins, Toby & Soler, Luc. (2021). U-Net Transformer: Self and Cross Attention for Medical Image Segmentation. 10.1007/978-3-030-87589-3_28.

[4] Zhou, Zongwei & Rahman Siddiquee, Md Mahfuzur & Tajbakhsh, Nima & Liang, Jianming. (2018). UNet++: A Nested U-Net Architecture for Medical Image Segmentation: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings. 10.1007/978-3-030-00889-5_1.

[5] [UW-Madison]. ([2022, July]).[UW-Madison GI Tract Image Segmentation]. Retrived from July, 2022 from

https://www.kaggle.com/competitions/uw-madison-gi-tract-image-segmentation/data

[6] Li, Yanghao & Wang, Naiyan & Liu, Jiaying & Hou, Xiaodi. (2017). Demystifying Neural Style Transfer. 2230-2236. 10.24963/ijcai.2017/310.

[7] Shamshad, Fahad & Khan, Salman & Zamir, Syed Waqas & Khan, Muhammad Haris & Hayat, Munawar & Khan, Fahad & Fu, Huazhu. (2022). Transformers in Medical Imaging: A Survey.

[8] Liu, Xiangbin & Song, Liping & Liu, Shuai & Zhang, Yudong. (2021). A Review of Deep-Learning-Based Medical Image Segmentation Methods. Sustainability. 13. 1224. 10.3390/su13031224.

[9] Dosovitskiy, Alexey & Beyer, Lucas & Kolesnikov, Alexander & Weissenborn, Dirk & Zhai, Xiaohua & Unterthiner, Thomas & Dehghani, Mostafa & Minderer, Matthias & Heigold, Georg & Gelly, Sylvain & Uszkoreit, Jakob & Houlsby, Neil. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

[10] Liu, Ze & Lin, Yutong & Cao, Yue & Hu, Han & Wei, Yixuan & Zhang, Zheng & Lin, Stephen & Guo, Baining. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows.

[11] Zhang, Yundong & Liu, Huiye & Hu, Qiang. (2021). TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. 10.1007/978-3-030-87193-2_2.

[12] Huang, Huimin & Lin, Lanfen & Tong, Ruofeng & Hu, Hongjie & Qiaowei, Zhang & Iwamoto, Yutaro & Han, Xian-Hua & Chen, Yen-Wei

[17] Touvron, Hugo & Cord, Matthieu & Douze, Matthijs & Massa, Francisco & Sablayrolles, Alexandre & Jégou, Hervé. (2020). Training data-efficient image transformers & distillation through attention.

& Wu, Jian. (2020). UNet 3+: A Full-Scale Connected UNet for Medical Image Segmentation.s

[13] Najeeb, Suhail & Bhuiyan, M.. (2022). Spatial feature fusion in 3D convolutional autoencoders for lung tumor segmentation from 3D CT images. Biomedical Signal Processing and Control. 78. 103996. 10.1016/j.bspc.2022.103996..

[14] Meng, Zhu & Fan, Zhongyue & Zhao, Zhicheng & Su, Fei. (2018). ENS-Unet: End-to-End Noise Suppression U-Net for Brain Tumor Segmentation. Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2018. 5886-5889. 10.1109/EMBC.2018.8513676.

[15] Wu, Junyan & Chen, Eric & Rong, Ruichen & Li, Xiaoxiao & Xu, Dong & Jiang, Hongda. (2019). Skin Lesion Segmentation with C-UNet. Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference. 2019. 2785-2788. 10.1109/EMBC.2019.8857773.

[16] H. Kan et al., "ITUnet: Integration Of Transformers And Unet For Organs-At-Risk Segmentation," 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Glasgow, Scotland, United Kingdom, 2022, pp. 2123-2127, doi: 10.1109/EMBC48229.2022.9871945.

[18] Yang, Jianwei et al. “Focal Self-attention for Local-Global Interactions in Vision Transformers.” *ArXiv* abs/2107.00641 (2021): n. pag.