Analysis of TransUNet: Hybrid CNN-Transformer

Kutay Ugurlu

Contents

I Introduction 3

II Theory 3
II-A U-Net 3
II-B Attention 4
II-C Extending Attention to Vision 4
II-D TransUNet 5
II-D1 Hybrid Model 5
II-D2 Cascaded Upsampler (CUP) 5

III Experiments and Results 5
III-A Performance Metrics 5
III-B Original Results [4] 6
III-C Replicated Results 7
III-C1 Training 7
III-C2 Inference 8

IV Conclusion 8

List of Figures

1 Whole Heart Segmentation [6] 3
2 Whole Brain Segmentation [7] 3
3 A generic U-Net Architecture 3
4 Words' contextual relation in the sentence [9] 4
5 MultiHead Self Attention Mechanism 4
6 Vision Transformer 4
7 TransUnet Framework 5
8 Comparsion on SYNAPSE dataset 6
9 Qualitative results for SYNAPSE dataset 6
10 False positive predictions in Slice 173 in concatenated data 7
11 Predictions in Slice 61 in concatenated data 7
12 Cross entropy loss during training 7

List of Tables

I Automated Cardiac Diagnosis Challenge dataset results 6
II Segmentation performance on individual organs utilized in the study 8

Analysis of TransUNet: Hybrid CNN-Transformer

Kutay Ugurlu

***Abstract*—Medical image segmentation is a crucial part of developing healthcare systems, not only in imaging in the conventional sense but also in electrical source imaging of the brain and heart [1]. So far, U-Net[2, text] has become ubiquitous in medical image segmentation tasks due to its recognized performance. However, the intrinsic features of convolutional neural networks(CNNs) that extract the local features of the spatial region in the image, the architecture exhibits some limitations in finding the long-range patterns. On the other hand, transformers, that were developed for sequence-to-sequence predictions, can recognize the global patterns via the utilization of attention mechanism [3]. Chen, *et al.*, propose a hybrid CNN-Transformer network called TransUNet in their study [4] to tackle the problem of medical image segmentation exploiting both local and global dependencies in the images. In this project, the aforesaid study is analyzed and a series of experiments is conducted whose results are available online at the forked repository: https://github.com/kutay-ugurlu/TransUNet__Analysis.**

I. Introduction

Image segmentation is a semantic segmentation task where each of the pixels (or voxels) is assigned a class corresponding to the object that they represent. Medical image segmentation has been of particular interest to researchers for diagnosis, treatment, and even prognosis, in different fields. Given the small size of medical datasets available, the segmentation is a challenging task [5]. Furthermore, especially in human imaging, the segmentation boundaries need to be precise to make reliable decisions on the data. Hence, a model has to be trained to separate the organ and physiological boundaries successfully with a high spatial resolution, as in Figure 2, while also having the capability to generalize to other data distributions it may be fed.

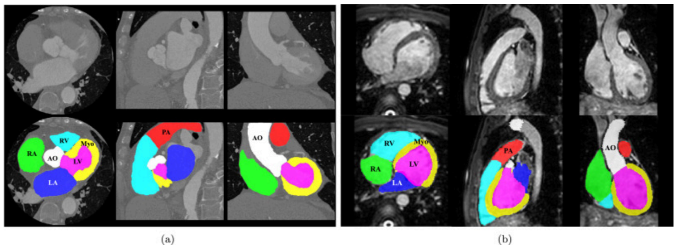
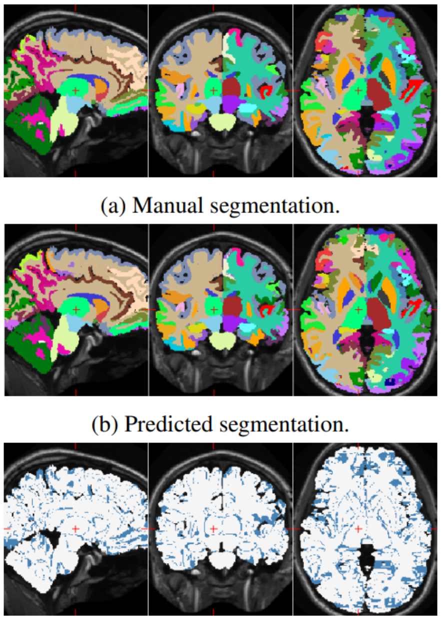


Fig. 1: Whole Heart Segmentation [6]

To improve both the local and global pattern recognition capability, the authors of [4] propose a hybrid network whose components are going to be analyzed individually in the following sections.



(a) Manual segmentation.

(b) Predicted segmentation.

(c) Difference, white voxels are identical while blue are different.

Fig. 2: Whole Brain Segmentation [7]

II. Theory

A. U-Net

U-Net is a convolutional neural network that utilizes certain modifications compared to its predecessors.

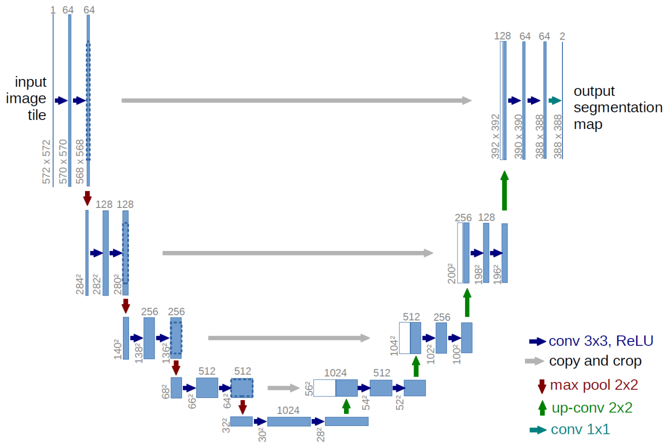


Fig. 3: A generic U-Net Architecture

The U-Net Architecture has the following properties that enable it to perform well in assigning classes to each pixel:

- Multilevel Decomposition: For shrinking image sizes, it utilizes the same size of convolutional kernels. In other words, for convolution, it employs different Fourier transform sizes to match different images size allowing it to decompose the image in multiple different spatial frequencies.
- Residual Learning: The skip connections created between input and output, along with the encoded features in the encoder pathway and the decoded image from latent features, helps the model to learn also from the features encoded before the latent representation and speeds up the training and provide a good convergence rate, avoiding the vanishing gradient problem, borrowing the idea that was proposed for ResNet in [8].

Thanks to the multichannel filtering feature of the CNNs, U-Net outputs usually more than one channel corresponding to the foreground (with the segmentation class counts) and the background. Using a proper cost function, such as mean squared error and an appropriate optimizer, the model learns the function that maps the pixels to the classification labels.

B. *Attention*

Attention was first proposed for sequence-to-sequence tasks, such as machine translation. Since it is developed for natural language processing, its mechanism is illustrated using ”words”. In tasks, such as machine translation, the input's contextual location or the other words in the sentence might have a huge effect on the input, as in Figure 4, where the word ”it” refers to two different words although it is in the same location in the sentence. However, what it refers to is changed by global dependency. In other words, the word at the very end of the sentence had an impact on the word itself. Hence, contextual similarity should also be learned when training sequence-to-sequence models.

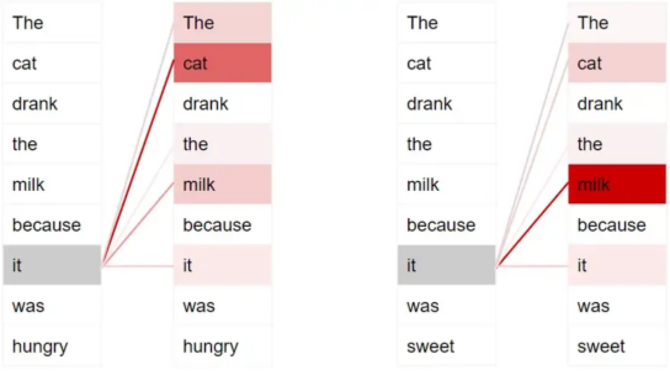


Fig. 4: Words' contextual relation in the sentence [9]

To learn such a relationship, the following mechanism is proposed:

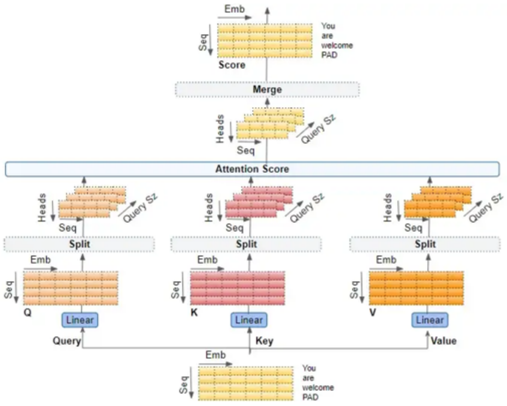


Fig. 5: MultiHead Self Attention Mechanism

In Figure 5, the sentence ”You are welcome.” is padded by one word, and the words are transformed into vectors, or embedding, via an embedding network such as Word2Vec [10]. In this example, the embedding size is 6. Then, the embedding is transformed by some trainable linear transformations, or matrices W_Q, W_K, W_V, to three different spaces, **Query**, **Key** and **Value**. Following this operation, the projections are split into different attention heads, and the relationship weights are calculated using $\sigma(QK^T)$ where σ is softmax. The weights corresponding to the `PAD` values are masked out and the resultant matrix is multiplied with matrix $V = W_v E$ where E is the input embeddings to obtain the attention score. The multi-head attention splits the transformed embedding and provides an opportunity for a different part of embedding to learn different relations. For instance, one part of the embedding can encode the gender(in languages with gender pronouns) and another part of the embedding can encode the cardinality(singular or plural).

C. *Extending Attention to Vision*

One of the main works whose results were utilized in the study is Vision Transformer proposed in [11].

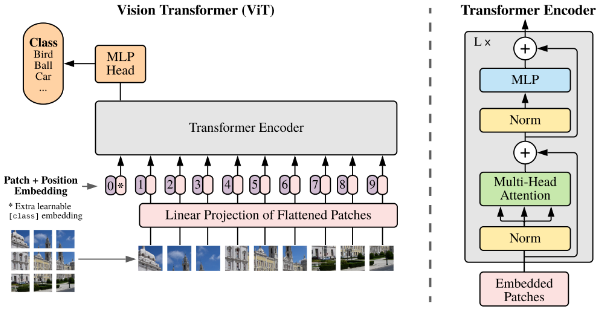


Fig. 6: Vision Transformer

In the overview of this model, we observe that the image is split into fixed-size patches and is projected linearly

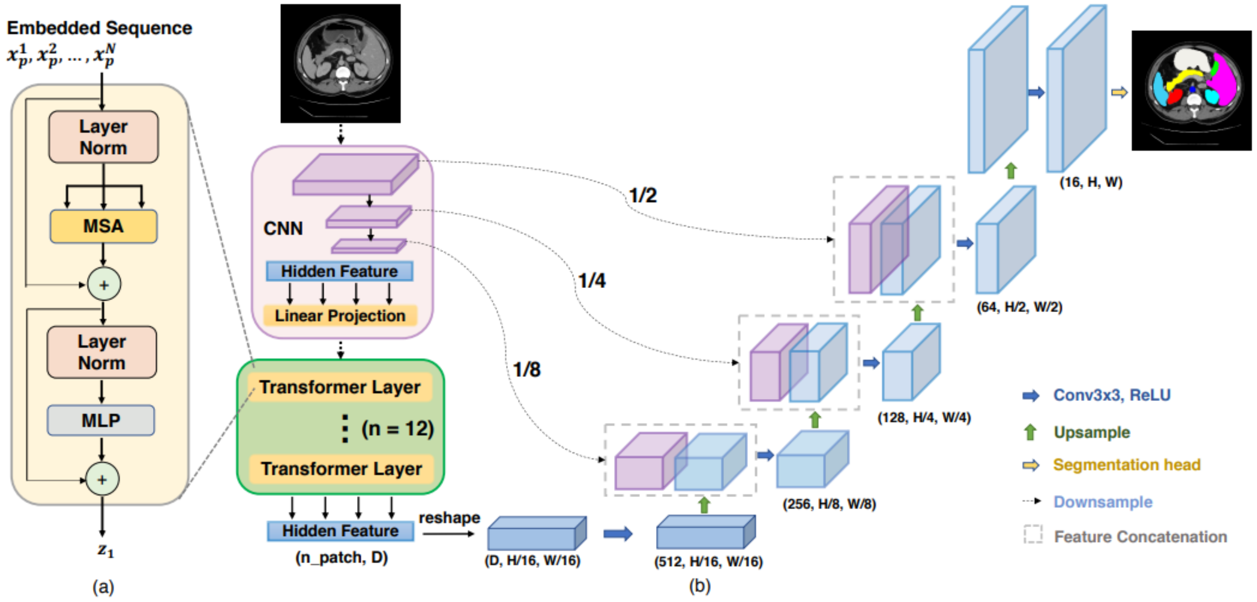


Fig. 7: TransUnet Framework

after flattening, with an extra positional embedding. After linear projections, these embeddings are fed into a transformer encoder which includes batch normalization, multi-head attention, a multilayer perceptron, and 2 skip-connections. By replacing the word embeddings with flattened image patches, Dosovitskiy *et al.* was able to perform classification on the images.

D. TransUNet

1) Hybrid Model

TransUNet is a hybrid network that exploits the local feature encoding characteristics of CNNs and global feature encoding characteristics of the transformers at the same time. As a modification on [11], TransUNet employs a CNN as a feature extractor to generate a feature map. Then, patch embedding is applied to 1 × 1 patches extracted from the CNN feature map instead of the raw images.

The authors list 2 reasons for following this approach: 1) it allows the utilization of the CNN feature map in the decoding path; 2) it results in better performance when compared to the case the inputs are taken as raw images.

$$z_0 = [x_p^1 E; x_p^2 E; \ldots; x_p^N E] + E_{pos} \tag{1}$$

where E is the patch embedding projection and E_{pos} is the positional embedding. Once the input of the first transformer layer is obtained as described in Equation (1), the l-th layer output of the transformer can be written as follows:

$$z'_l = MSA(LN(z_{l-1})) + z_{l-1} \tag{2}$$

$$z_l = MLP(LN(z'_l)) + z'_l \tag{3}$$

2) Cascaded Upsampler (CUP)

The decoded feature representation $z_L \in \mathbb{R}^{\frac{HW}{P^2}} \times D$ is reshaped to $\frac{H}{P} \times \frac{W}{P}$. Then, by using 1×1 convolution, the number of channels is reduced to the number of classes. This approach directly applies bilinear upsampling (interpolation) on the coded representation and the resultant tensor is the output result. This naïve interpolation approach is called ”None” in [4]. Instead of this, the researchers cascade multiple upsampling blocks consisting of 2× upsampling operator, a 3×3 convolution layer, and ReLU nonlinearity successively.

III. Experiments and Results

A. Performance Metrics

The metrics defined in Equations (4) and (5) are utilized for the evaluation of the performance.

$$Dice\ Coeff.(DSC) = 2 \times \frac{A \cap B}{A \cup B} \tag{4}$$

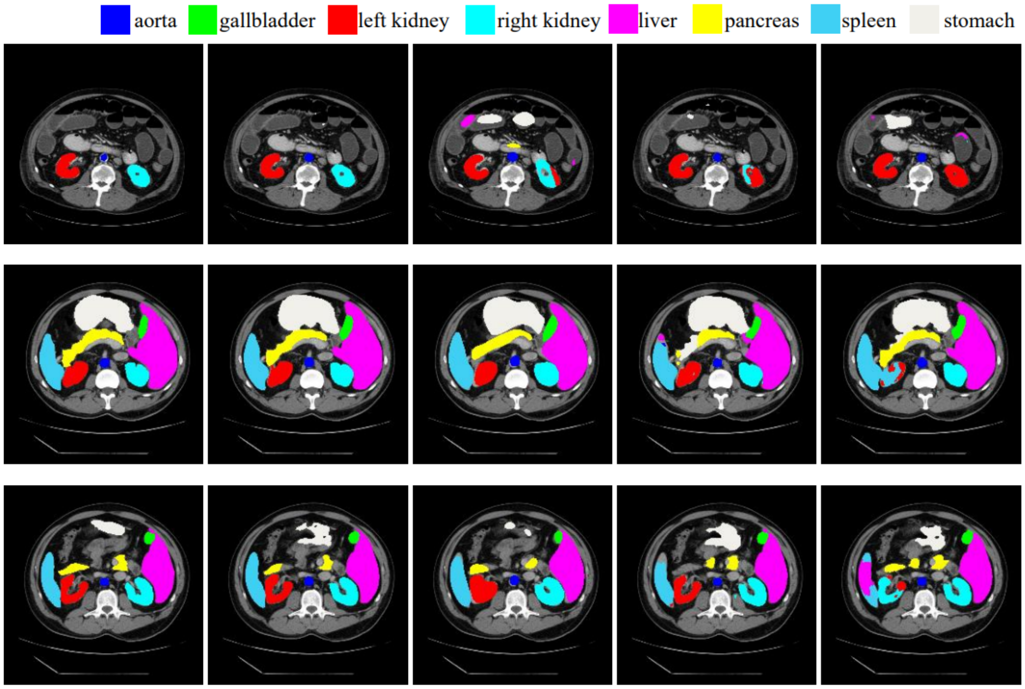
$$d_H(X,Y)(HD) = \max\left\{\sup_{x\in X} d(x,Y), \sup_{b\in B} d(X,y)\right\} \tag{5}$$

where A and B in Equation (4) are target and found segmentation regions, whereas X and Y in Equation (5), the Hausdorff distance, are the segmentation boundaries corresponding to the segmented regions.

When the model predictions show an exact match with labeled regions, both metrics are 1.

Framework		Average		Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas	Spleen	Stomach
Encoder	Decoder	DSC ↑	HD ↓								
V-Net [9]		68.81	-	75.34	51.87	77.10	80.75	87.84	40.05	80.56	56.98
DARR [5]		69.77	-	74.74	53.77	72.31	73.24	94.08	54.18	89.90	45.96
R50	U-Net [12]	74.68	36.87	84.18	62.84	79.19	71.29	93.35	48.23	84.41	73.92
R50	AttnUNet [13]	75.57	36.97	55.92	63.91	79.20	72.71	93.56	49.37	87.19	74.95
ViT [4]	None	61.50	39.61	44.38	39.59	67.46	62.94	89.21	43.14	75.45	69.78
ViT [4]	CUP	67.86	36.11	70.19	45.10	74.70	67.40	91.32	42.00	81.75	70.44
R50-ViT [4]	CUP	71.29	32.87	73.73	55.13	75.80	72.20	91.51	45.99	81.99	73.95
TransUNet		**77.48**	**31.69**	87.23	63.13	81.87	77.02	94.08	55.86	85.08	75.62

Fig. 8: Comparsion on SYNAPSE dataset



(a) GroundTruth **(b) TransUNet** **(c) R50-ViT-CUP** **(d) AttnUNet** **(e) UNet**

Fig. 9: Qualitative results for SYNAPSE dataset

Framework	Average	RV	Myo	LV
R50-U-Net	87.55	87.10	80.63	94.62
R50-AttnUNet	86.75	87.58	79.20	93.47
ViT-CUP	81.45	81.46	70.71	92.18
R50-ViT-CUP	87.57	86.07	81.88	94.75
TransUNet	89.71	88.86	84.53	95.73

TABLE I: Automated Cardiac Diagnosis Challenge dataset results

B. *Original Results [4]*

There are a series of ablation studies to determine the number of skip connections, input resolution, patch size, and sequence length. The latter three are dependent on each other by the relation $P^2L = WH = W^2$, where P is the patch size, L is the sequence length, and W and H are the input image sizes. PyTorch [12] deep learning framework is used to train the models. The models are trained on a single Nvidia RTX2080Ti GPU with the following hyperparameters:

- Input size: 224 × 224
- Patch size: 16
- Learning rate: 0.01
- Momentum 0.9
- Weight decay: 0.0001
- Optimizer: Stochastic Gradient Descent
- Number of training iterations: 20000
- Default batch size: 24

The TransUNet model that is used to produce these results is what is referred to as **base** model in the ablation results. To train the proposed network, Chen *et al.* used 30 abdominal CT scans in the MICCAI 2015 Multi-Atlas Abdomen Labelling Challenge, adding up to 3779 axial

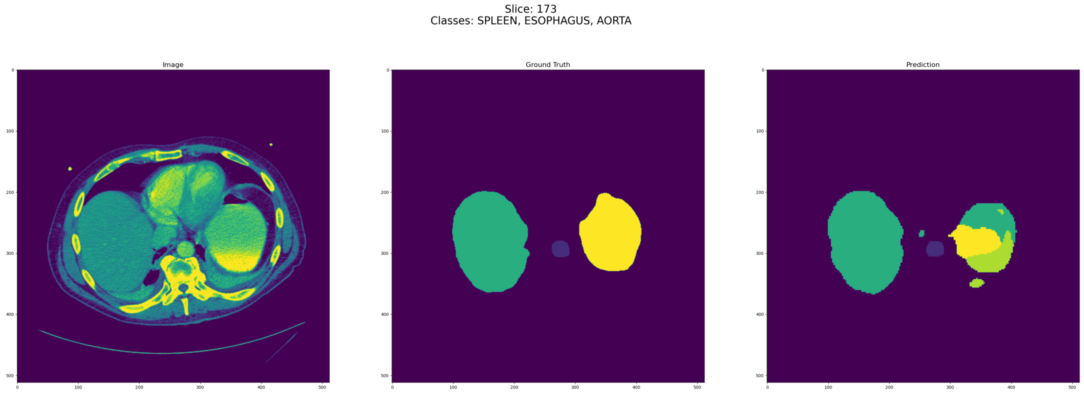


Fig. 10: False positive predictions in Slice 173 in concatenated data

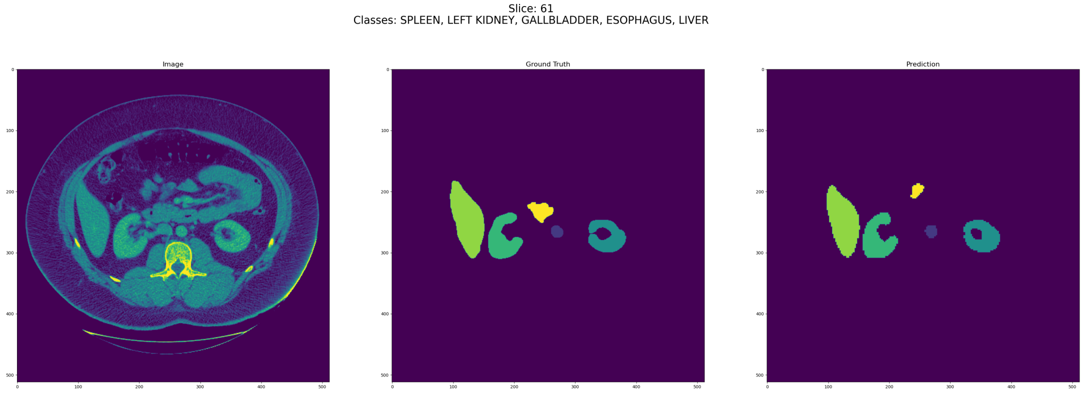


Fig. 11: Predictions in Slice 61 in concatenated data

contrast-enhanced images from the SYNAPSE dataset and ACDC dataset. The experiments are done with different encoder and decoder architectures and the average DSC in percentage and Hausdorff distance in millimeters for eight abdominal organs with a random split 18 training and 12 validation cases are reported. Looking at Figure 8, we see that TransUNet outperforms the best predecessor by 2.8% in terms of average DSC and it reduced the HD by approximately 3% when compared with the best predecessor. It is possible to investigate that the proposed method's performance is the best for 4 organs presented in Figure 8. In Table I, we observe that for both metrics, TransUNet outperformed the others for segmentation classes Left Ventricle, Myocardium, and Right Ventricle.

C. *Replicated Results*

1) *Training*

The models are trained on a single NVIDIA GeForce GTX 1660 Ti on TU116-A GPU with the following hyperparameters:

- Input size: 224 × 224
- Patch size: 16
- Learning rate: 0.01
- Momentum 0.9
- Weight decay: 0.0001
- Optimizer: Stochastic Gradient Descent
- Number of training epochs: 150
- Batch size: 8

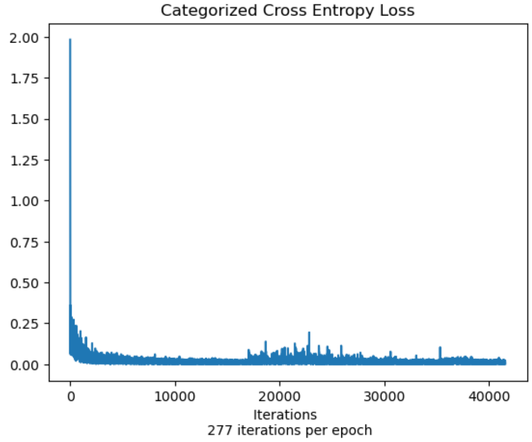


Fig. 12: Cross entropy loss during training

For training, argparser that is provided in the original repository is utilized as explained the documentation. The batch size is updated due to the limited GPU memory size. With the given configuration above, the training took 18 hours with 150 epochs on 83 scans with 1009 slices that were provided as averaged training images.

2) Inference

On inference time, 12 different test cases were evaluated. The inference took approximately 2 minutes per case on average. DSC and HD are calculated between the ground truth and the model outputs slice-by-slice and the mean of the metrics along the slice dimension is calculated for each test case. Then, mean±std DC metric is calculated along with the median(IQR) HD metric are calculated for each test case:

- DSC: 0.7810 ± 0.0841
- HD: 10.9264 (35.5377)

The network produces only 8 organs' individual results. For the inference whose results given above, the corresponding evaluations for individual organs are presented in Table II. Although most of the predictions match with corresponding ground truth labelings as shown in Figure 11, there were some cases where some false positive detections are identified. This may be due to the undertrained model due to the given limited time and computational resources. Although we expect the model to find 3 different classes in Figure 10, there were 5 different classes predicted in the inference time.

Metric / Organ	DSC	HD (mm)
Aorta	0.86	13.85
Gallbladder	0.64	18.74
Spleen	0.83	41.36
Left Kidney	0.79	33.65
Right Kidney	0.92	37.61
Liver	0.61	15.07
Pancreas	0.86	45.46
Spleen	0.76	21.03

TABLE II: Segmentation performance on individual organs utilized in the study

IV. Conclusion

In this project report, the study *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation* is analyzed and some of the results were replicated in different training and testing configurations. The results are satisfying given the model's size and predictions, even when compared with the state-of-the-art. Although detailed ablation studies for different hyperparameters are already conducted, the inference can be improved by training the model on a dataset that has more diverse distributions.

References

[1] A. González-Ascaso, R. Molero, A. M. Climent, and M. S. Guillem, "Ecgi metrics in atrial fibrillation dependency on epicardium segmentation," in *2020 Computing in Cardiology*, IEEE, 2020, pp. 1–4.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, Springer, 2015, pp. 234–241.

[3] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] J. Chen, Y. Lu, Q. Yu, *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[5] X. Ren, L. Zhang, D. Wei, D. Shen, and Q. Wang, "Brain mr image segmentation in small dataset with adversarial defense and task reorganization," in *International Workshop on Machine Learning in Medical Imaging*, Springer, 2019, pp. 1–8.

[6] X. Zhuang, L. Li, C. Payer, *et al.*, "Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge," *Medical Image Analysis*, vol. 58, p. 101 537, 2019, ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2019.101537. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841519300751.

[7] A. de Brebisson and G. Montana, "Deep neural networks for anatomical brain segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 20–28.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[9] K. Doshi, *Transformers explained visually (part 3): Multi-head attention, deep dive*, Jun. 2021. [Online]. Available: https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853.

[10] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[12] A. Paszke, S. Gross, S. Chintala, *et al.*, "Automatic differentiation in pytorch," 2017.