DOI: 10.37943/13BKBF2003

Nam Diana
Master of Tech. Sci., PhD Student
di_nam@kbtu.kz, orcid.org/0000-0002-9356-3114
School of Information Technology and Engineering,
Kazakh British Technical University, Almaty, Kazakhstan

Pak Alexandr Alexandrovich
Candidate of tech. sciences, Professor
a.pak@kbtu.kz, orcid.org/0000-0002-8685-9355
School of Information Technology and Engineering,
Kazakh British Technical University, Almaty, Kazakhstan

OVERVIEW OF TRANSFORMER-BASED MODELS FOR MEDICAL IMAGE SEGMENTATION

Abstract: Premedical diagnostics is the process of examining survey results. Correct premedical diagnostics can improve the process of patient management and reduce the burden on the medical sector. Diagnostics of medical images such as computed tomography and X-ray are an obligatory step for further treatment. However, the shortage of clinicians causes delays in this step. We observed two state-of-the-art algorithms proposed for medical image segmentation: TransUnet and Swin-Unet. We conducted a theoretical comparison of algorithms in terms of the applicability of pre-hospital diagnostics according to quality and speed of training. The comparison is based on the original source of code provided by the authors of the original articles. We chose these two algorithms because they have similar U-form architecture, a high level of citation, and show competitive DICE scores on pictures of various human organs. Some architectural features were also important. Both models inherit key elements of U-net. TransUnet is a hybrid Transformer and CNN model. It consists of Transformer encoder and a convolutional decoder. Some additional computations are required in the bottleneck. Swin-Unet is a fully Transformer-based model. These architectural differences give rise to a difference in the number of trainable parameters. Generally, deeper architectures with bigger number of parameters usually show better performance, however, according to our review, Swin-Unet has smaller number of parameters and shows better DICE and Hausdorff Distance. It should be noted that the distribution between false positive and false negative predictions is important in medical image processing. It is crucial to avoid overloading the medical sector while also not missing any sick patients. Precision and recall can be used to evaluate the ratio of incorrect predictions. Therefore, we also observed the results of caries segmentation where precision and DICE were provided. In this specific case, TransUnet shows better DICE and recall values but worse precision.

Keywords: Computer Vision, Transformers, Image processing, premedical diagnostics, Segmentation

Introduction

In the modern world, there is a big problem with a lack of clinicians. It causes a problem with patient management. Patients with a deeper stage of the disease cannot receive timely help. This is especially acute in agrarian countries because the examination time is delayed due to communication difficulties between the regions. For example, the ratio between the rural and urban populations of Kazakhstan is 54.1 to 45.9%. The shortage of doctors in the country has doubled in the last five years. In this particular case, the application of computer technologies can reduce the burden on the medical sector. It could be applied for premedical diagnostics task.

In the current review, we observed and compared two algorithms proposed for the medical image segmentation task. Image segmentation is the process of derivation of each pixel of the image into categories. It is the derivation between healthy and diseased cells for medicine.

Convolutional neural networks [1] are the most commonly used algorithms for medical image processing. Common highly accurate neural networks, such as AlexNet [2] and ResNet [4], demonstrate high accuracy in solving tasks related to medicine as well as for general use cases. However, it has an important architectural disadvantage. The vision of the network is limited by the size of the sliding window. It makes it impossible to view the whole picture at one point in time. Initially, the usage of "sliding window" was justified by the lack of computing power. The further rapid growth of computing systems allowed the use of more advanced models for computer vision [8] [9] [10].

We stopped on two Transformer-based models proposed for the medical image segmentation task: TransUnet [9] and Swin-Unet [8]. Both architectures are based on the U-net model [7], which was proposed in 2019 for medical image segmentation. U-net is a convolutional-based model with three key features: high-quality processing of medical images. Firstly, it could be trained end-to-end on the small dataset. It is essential because of the lack of medical data for rare diseases and the confidentiality of medical data. Secondly, it takes into account pixels near the border of the image. Last but not least, it allows you to segment thin borders on the image, such as borders between cells on a biopsy. These benefits are achieved by the mirror "U" form architecture and inherits by TransUnet and Swin-Unet models. The convolutional encoder was replaced by Vision Transformer (ViT) [6] in TransUnet which allows to increase DICE in medical image segmentation. A combination of Swin Transformer [12] and U-Net formed Swin-Unet. The authors of Swin-Unet architecture provided a model comparison among DICE, Hausdorff Distance on Synapse multiorgan computer tomography dataset and among DICE on ACDC dataset. We completed the comparison by computing the number of training parameters for both models based on code listing provided by the authors of the original models.

Transformer models for computer vision tasks

The TransUNet model inherits key advantages of the U-net algorithm with the addition of the transformer's features [9]. ViT transformers [6] proposed for classification task encoder were completed by MLP for solving a classification task. TransUnet consists of transformer encoder, convolutional based decoder and skip connections between their corresponding layers. While U-net [7] was designed for end-to-end training on a small number of samples, transformers are pretrained on a large dataset, such as ImageNet [3]. The TransUnet algorithm is presented for solving medical image segmentation tasks.

The architecture of the model is shown of Fig. 1. In the U-net model, the encoder was shown as a usual convolutional neural network with two double convolutional layers and a max pooling operation after that. In TransUnet, it is replaced by a transformer block. The components of Transformers layers are provided in Fig. 1 (left).

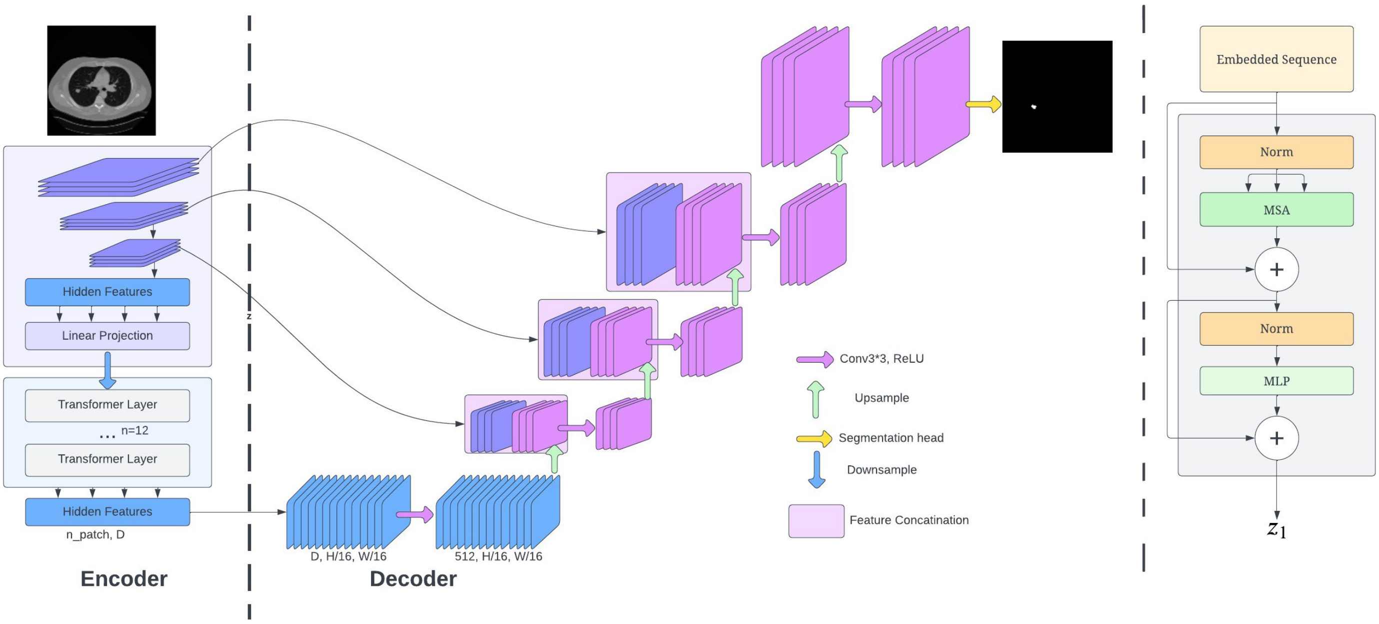


Figure 1. TransUnet model architecture and Transformer block[9]

Self-attention mechanisms based on transformers are used for feature extraction instead of usual CNN based approaches. Firstly input x is split into 2D patches $\{x_p^i \in \mathbb{R}^{P^2} * C | i = 1, \dots, N\}$ where P is a patch size. Trainable linear projection is used for mapping x into D-dimensional. Specific position embeddings are trained for the patch spatial information (1).

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

where $E \in \mathbb{R}^{p^{2*C}}$ is patch embedding projection

L layers of Multihead Self-Attention and Multi-Layer Perceptron (MLP) formed the encoder.

TransUnet was compared with Unet and ViT models on the Automated Cardiac Diagnosis Challenge (ACDC) [11] and showed better DICE (87.55, 87.57, 89.71 for -U-Net, ViT-CUP and TransUNet respectively).

While the decoder of the TransUnet model is a convolutional neural network with an upsampling operation inside, the decoder and encoder of the next observed model [8] are both transformer-based.

Swin-Unet The second observed model is Swin U-net [8] proposed in 2021. It also solves medical image segmentation task. Swin U-net inherits similar structure with U-net [7] and Swin Transformers block [12]. It has an encoder, decoder, skip connections and bottleneck. The CNN-based encoder was replaced by a Transformer block in Trans U-net architecture [9]. Swin U-net contains Transformer blocks inside encoder and decoder both. The architecture of the model is provided in Fig 2.

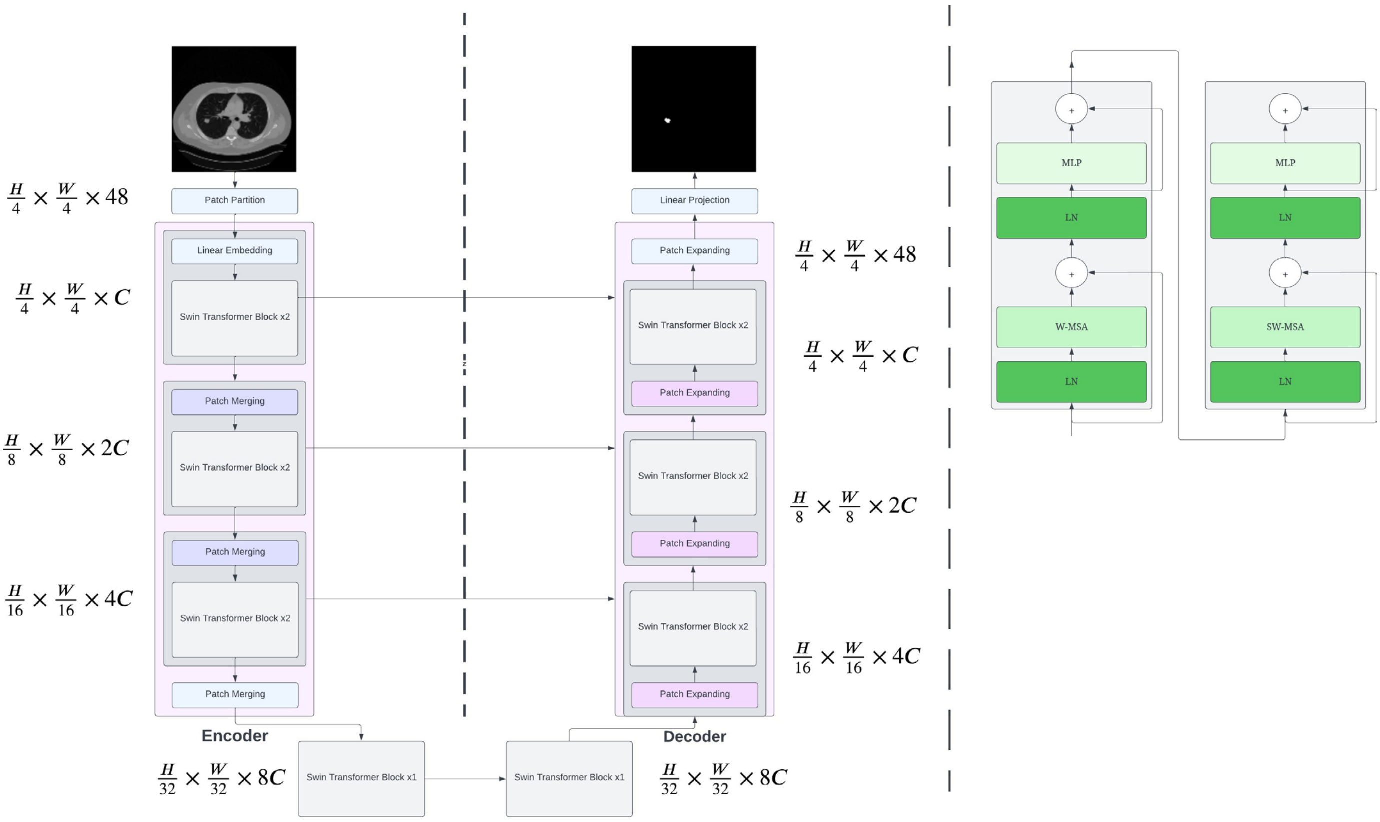

$\frac{H}{4} \times \frac{W}{4} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
Patch Partition
Linear Embedding
Swin Transformer Block x2
Patch Merging
Swin Transformer Block x2
Patch Merging
Swin Transformer Block x2
Patch Merging
Encoder
Swin Transformer Block x1
Swin Transformer Block x1
Decoder
Linear Projection
Patch Expanding
Swin Transformer Block x2
Patch Expanding
Swin Transformer Block x2
Patch Expanding
Swin Transformer Block x2
Patch Expanding
$\frac{H}{4} \times \frac{W}{4} \times 48$
$\frac{H}{4} \times \frac{W}{4} \times C$
$\frac{H}{8} \times \frac{W}{8} \times 2C$
$\frac{H}{16} \times \frac{W}{16} \times 4C$
$\frac{H}{32} \times \frac{W}{32} \times 8C$
MLP
LN
W-MSA
LN
MLP
LN
SW-MSA
LN

Figure 2. Swin-Unet model architecture [8]

The encoder block is provided as a transformer with a sequence embedding input. So the input image should be split into non-overlapping patches with the size 4 * 4. Because of this operation, the feature dimension of each patch is 4 * 4 * 3 = 48 (where 3 is the number of channels). C on Fig. 2 is a linear embedding layer. It was applied to the dimension of a projected feature in an arbitrary dimension. The hierarchical feature representations are generated by Swin Transformer blocks and patch merging layers. Down-sampling and increasing dimension obtained through patch merging layer and feature representation learning achieved through the usage of Swin Transformer block. Swin Transformer block was described as a multipurpose backbone for Transformer-based model and tested for both classification and object detection [12]. It is a key element of Swin U-net architecture.

While in computer vision tasks convolutional neural networks are the predominant type of architecture, neural network-based algorithms seriously lose to Transformer-based algorithms in natural language tasks [5]. The creation of transformers for sequence data allows them to process data with long-range dependencies. According to the authors of Swin transformer [12], there are two main challenges in adapting transformers for computer vision. The first one is identifying a basic element. While the word tokens are the basic element for the language processing transformer, it is impossible to identify the fixed size of the minimum element of the image for computer vision Transformer. The second problem is the computation complexity, which is caused by the higher number of pixels per image than a word in a paragraph [12]. According to the authors of the Swin U-net architecture, their model could solve both of them by Hierarchical feature maps and shifted window approaches. Swin Transformer architecture is shown in Fig. 2.

As shown in Fig. 2, the input image with 3 channels is split into non-overlapping patches. In the implementation provided in the original article, the size of each patch is 4 * 4 * 3. So the total number of patches is (H * W).

While in previous observed articles [6], [9], the image is split into patches with a fixed scale, Swin Transformer constructed hierarchical feature maps as provided in Fig 3. This approach imitates the feature map resolutions of the usual convolutional neural network [13] [14]. The number of patches decreases in the deep layers, by concatenation with the neighbor with $2 * 2$. Down-sampling resolution is 2 times, and the output dimension is 2.

The representation in hierarchical feature maps in Swin transformers starts from small size patches colored by grey in Fig 3 (a) which gradually merge with neighbor patches in deep layers. The usage of hierarchical feature maps may easily use sophisticated dense prediction algorithms like feature pyramid networks (FPN) [15] and U-net [7]. As it is shown in Fig 3 (b) the image is split into fixed $16 * 16$ size patches in Vision transformer. It makes ViT unsuitable for cases when target object is much smaller than the image size, especially for semantic segmentation task, when the model should observe the belonging of each pixel.

Another problem with ViT transformers is their quadratic computational complexity. It is caused by the global self-attention mechanism which processes all patch vectors. But the number of tokens increases with the image size. Thereby, the authors of Swin Transformers proposed to calculate self-attention within a non-overlapping window (shifted window). So standard multi-head self-attention (MSA) is replaced by shifted windows (Fig 4) in the Swin Transformer block. It allows getting linear-complexity instead of quadratic in Swin Transformer. Global self-attention could be formalized via (2).

$$\Omega(MSA) = 4hwC^2 + 2(hw)^2C \tag{2}$$

where MSA is standard multi-head attention $[h, w]$ is image dimension, computation complexity of MSA is quadratic to the image size, C is the length of patch vector

While window-self attention has linear complexity, which could be formalized via (3)

$$\Omega(W - MSA) = 4hwC^2 + 2M^2(hw)C \tag{3}$$

where $W - MSA$ is shifted window-based MSA, M is the fixed window size, default $\mathrm{M} = 7$, the computation complexity is linear to the image size. The local window is colored red in Figure 4. A regular window partitioning strategy is used in the left part of Fig 4. It starts from the top-left pixel. The feature size $8 * 8$ is separated into windows with size $4 * 4$. Then the window slides from the previous layer by shifting the window by $\left(\left\lfloor\frac{M}{2}\right\rfloor, \left\lfloor\frac{M}{2}\right\rfloor\right)$ pixels from the regularly partitioned. The total number of patches should be saved. So the blocks are transposed with each other.

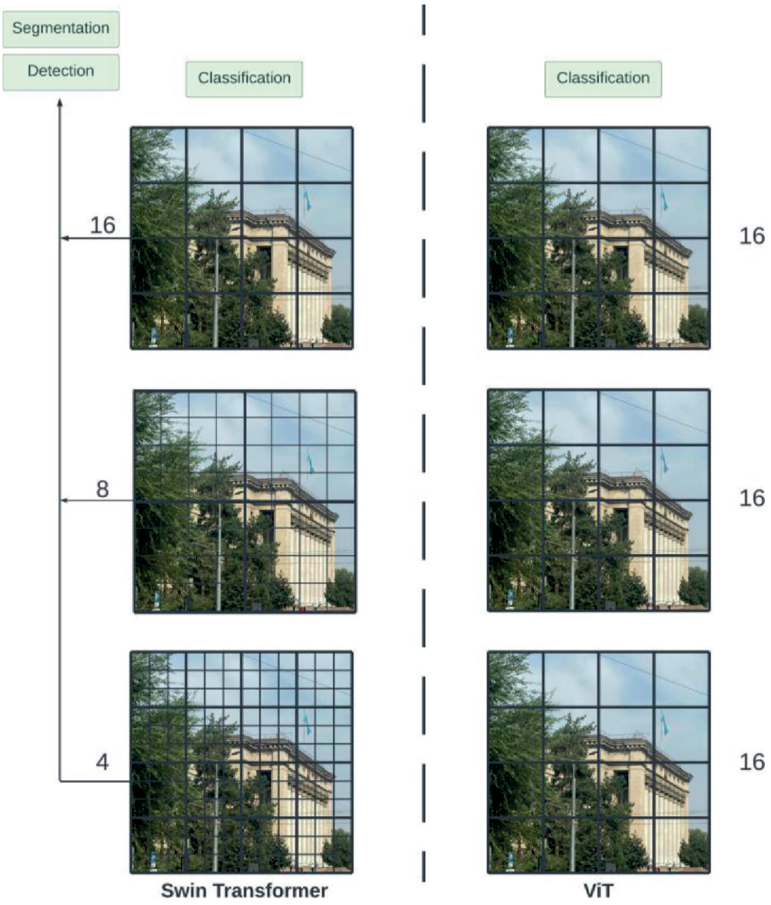


Figure 3. Hierarchical feature maps in Swin Transformer
(a) and fixed size in Vision Transformers ViT (b) [12]

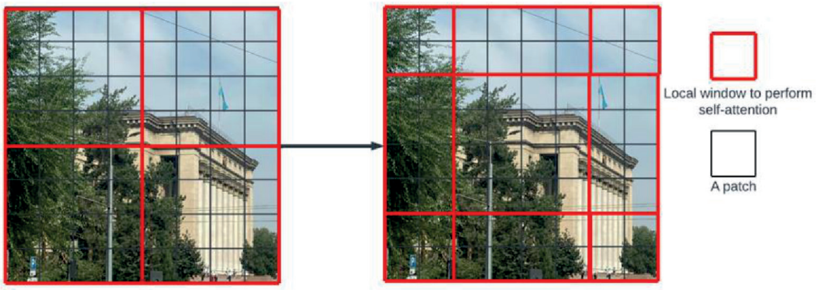


Figure 4. Shifted windows [12]

Swin Transformer block is calculated consecutively with the usage of shifted window-based multi-head self-attention (4-7).

$$\widehat{z^l} = W - MSA\left(LN(z^{l-1})\right) + z^{l-1}, \tag{4}$$

$$\widehat{z^l} = MLP\left(LN(\widehat{z^l})\right) + \widehat{z^l}, \tag{5}$$

$$\widehat{z^{l+1}} = SW - MSA\left(LN(z^l)\right) + z^l, \tag{6}$$

$$z^{\widehat{l+1}} = MLP\left(LN\left(z^{\widehat{l+1}}\right)\right) + z^{\widehat{l+1}}, \tag{7}$$

where $W-MSA$ is multi-head self-attention. $SW-MSA$ is Shifted Multi-head self attention, MLP is multi-layer perception, $\widehat{\mathrm{z}^{\mathrm{l}}}$ is an output the $(S)W-MSA$ and $\mathrm{z}^{\mathrm{l}-1}$ is an output of MLP, LN is the natural logarithm.

Relative position bias is used for calculation the distance between the patches. It is based relative position representations described in the article [20]. Self-attention mechanism is performed by (8).

$$Attention(Q,K,V) = SoftMax\left(\frac{QK}{\sqrt{d}} + B\right)V \tag{8}$$

where $\mathrm{B} \in \mathrm{R}^{\mathrm{M}^{2} * \mathrm{M}^{2}}$ is Bias $\mathrm{Q} \in$ is query, $\mathrm{K} \in$ is key, $\mathrm{V} \in$ is value, d is the dimension of query and key, M^{2} is the number of patches in a window.

All of the above formulated Swin transformer blocks shown in Fig. 2.

The Swin transformer has been tested for classification, segmentation, and object detection tasks. And shows high performance for all of them [8]. It is used as a backbone for the encoder and decoder, both in the Swin U-net architecture (Fig. 2).

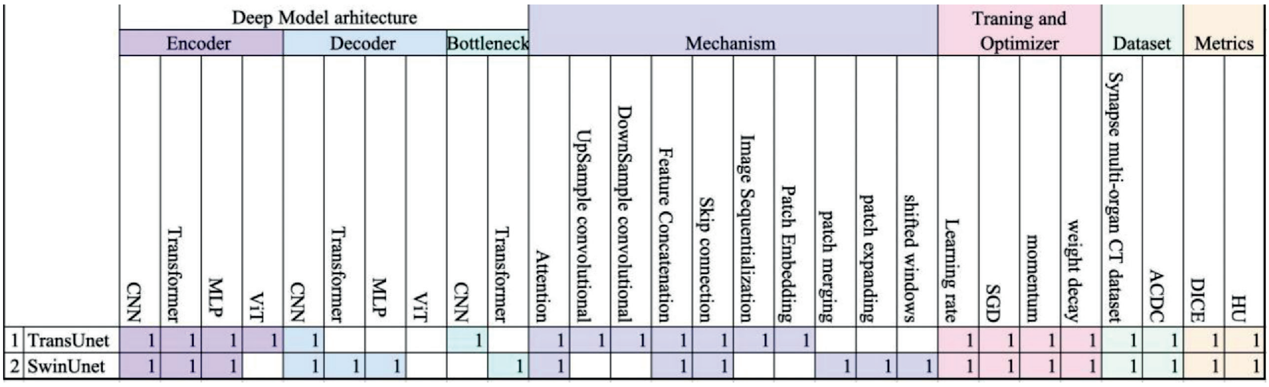
The decoder part of the model is presented as symmetric to the encoder transformer. It consists of a Swin Transformer block and a patch expanding layer. Skip connections are used for supplementing features that were lost during up-sampling by adding multiscale features from the encoder with the extracted context features. Expanding layer performed up-sampling in Swin-Unet model. The patch of encoder reshapes adjacent-dimension feature maps into big feature maps with 2 times up-sampling of resolution. The last patch is used for 4 times up-sampling for the reconstruction of the original size of the image. The output is fed to the linear projection layer to get the resulting image. The activation function is Gaussian error linear units (GELU).

The encoder consists of two sequential Swin Transformer blocks that complete representation learning. The feature dimension and resolution are saved during this operation. The number of patches decreases by 2 times and the feature dimension increases by 2 times on the patch merging layer. It is performed three times in the encoder.

Bottleneck of Swin-Unet consists of double Swin Transformer blocks. The resolution and feature dimension also is not changed in bottleneck.

Skip connection is used for the same goal as in the original U-net article. It sent features for the corresponding layers of the decoder. Small features are concatenated with the deep feature to decrease the number of relevant features lost because of down-sampling. The concatenated features have the same dimension as up-sampled features.

The comparison of TransUnet and Swin-Unet models is provided in Fig. 5. It shows deep model architecture, applied mechanism and optimizers used in both models. Architectures were compared on Synapse multi-organ computer tomography dataset and ACDC dataset based on DICE (9) and Housdorff Distance (10).



		Deep Model arhitecture										Mechanism										Traning and Optimizer				Dataset		Metrics	
		Encoder				Decoder				Bottleneck																			
		CNN	Transformer	MLP	ViT	CNN	Transformer	MLP	ViT	CNN	Transformer	Attention	UpSample convolutional	DownSample convolutional	Feature Concatenation	Skip connection	Image Sequentialization	Patch Embedding	patch merging	patch expanding	shifted windows	Learning rate	SGD	momentum	weight decay	Synapse multi-organ CT dataset	ACDC	DICE	HU
1	TransUnet	1	1	1	1	1				1		1	1	1	1	1	1	1				1	1	1	1	1	1	1	1
2	SwinUnet	1	1	1		1	1	1			1	1			1	1			1	1	1	1	1	1	1	1	1	1	1

Figure 5. Concept matrix

The authors provided a more in-depth comparison of segmentation SwinUnet and TransUnet on Medical Image Segmentation on the Synapse multi-organ computer tomography dataset [16]. The dataset consists of 30 computed tomography, with 85 – 198 slices of 8 abdominal organs: aorta, gallbladder, left kidney, liver, pancreas, right kidney, spleen, and stomach. The total number of slices is 3779. The data was split to 18 : 12 cases for training and validation, respectively. So the training set contains 2212 slices, and the validation set consists of 1567 slices. The original size of the image is $512 * 512$. It was resized to $224 * 224$ for further experiments.

The authors of the original articles evaluate the performance of the models by DICE (9) and Housdorff Distance (10).

$$DICE = \frac{2|X \cap Y|}{|X| + |Y|} \tag{9}$$

$$d_H(X, Y) = \max\left\{ \underset{x \in X}{sup}\ \underset{y \in Y}{inf} |xy|\, , \underset{y \in Y}{sup}\ \underset{x \in X}{inf} |xy| \right\} \tag{10}$$

Both metrics show the performance evaluation of image segmentation models by comparing ground true mask from the dataset and the output of the model. However, the cost of the model is formed not only on the quality of segmentation. The speed of model training and the required hardware should also be taken into account. The model parameters show how much data the model processes during training. The higher this number, the slower the model will run. It will also require more computing power. However, the motivation for the application of computer vision for medical image processing is to increase the quality of premedical diagnostics. This problem is more common in areas far from the city. For example, 80% of Kazakhstan is a rural place. In these regions, there is a severe shortage of qualified doctors and equipment. So, models with a smaller number of parameters are preferable.

The total number of parameters for the model is the sum of all trainable parameters in all network layers, including convolutional, transformer, normalization, and linear layers. All layers that form the deep model architecture are provided in Fig. 5. The authors of both models provide the code for their PyTorch implementation [17] and [18]. ResNet-50 was used as a convolutional backbone for TransUnet. CNN backbone was not required for the Swin-Unet model. The distribution of trainable parameters is 105276066 : 41376516 for TransUnet and Swin-Unet, respectively. So TransUnet requires 2.5 times more parameters than Swin-Unet to achieve competitive results. Also, Swin-Unet achieved a higher DICE and a smaller HU than TransUnet, which makes Swin-Unet more appropriate for usage because of economic and practical reasons.

Also, both models were compared with the original convolutional based Unet. DICE for Swin-Unet is 79.13, while for U-Net and TransUnet are 76.85 and 77.48, respectively on the Synapse multi-organ computer tomography dataset. So both observed transformer-based architectures show better performance for medical image segmentation than modern convolutional based architectures.

Table 1. TransUnet and Swin-Unet comparison

	TransUnet	Swin-Unet
Number of Transformer Layers	12	12
Input of img size	224*224	224*224
Convolutional backbone	ResNet-50	n/a
Input patch size	16	4
Pretrained on	ImageNet	ImageNet
Optimizer	SDG	SDG
LR	0.001	0.05
Momentum	0.9	0.9
Weight decay	1,00E-04	1,00E-04
Hardware	Nvidia RTX2080Ti GPU	Nvidia V100 GPU, 32 GB
Batch size	24	24
Max epoch	150	150
Total number of parameters	105276066	41376516
Average DICE	77.48	79.13
Average HU	31.69	21.55

Discussion

For medical image processing, we also need to take into account the distribution between True Positive and False Negative predicted pixels. As it was mentioned before, the medical image segmentation models could be applied as an intermediate step between computed tomography and doctoral examination. It will increase the quality of patient management. Cases, in which the model is identified as potentially affected by the disease, could be checked with higher priority by the specialist. Because it is premedical diagnostics and computed tomography images will be checked by the doctor one more time, we need to decrease the number of False Negative samples. Precision and Recall are used to evaluate the quality of the predicted mask among the distribution between True Positive and False Negative (11–12)

$$Precision = \frac{TP}{TP+FP} \qquad (11)$$

$$Recall = \frac{TP}{TP+FN} \qquad (12)$$

Dice, Precision and Recall for TransUnet and Swin-Unet based on caries segmentation on tooth X-ray images are provided on Table 2.

Precision is more important when False Negative costs more than False Positive. And vice versa, Recall is more important when False Positive costs more than False Negative. So, for the application of machine learning, we need to pay more attention to Recall. Because the authors of the original articles compared the models among DICE and HU only, we observed one more article [19] where DICE and Precision were provided. The authors proposed their model for caries segmentation on tooth X-ray images and compared it with U-net, TransUnet and Swin-Unet.

DOI: 10.37943/13BKBF2003
© Nam Diana, Pak Alexandr

The authors used 153 X-ray images, 40 of them were used for the validation set. The number of training images has been increased from 113 to 800 by standard augmentation methods, for example, rotation and translation. The input image size also was 224*224.

The authors provided DICE and Precision for U-net, TransUnet, Swin-Unet and their model. We calculated Recall from provided data and compared TransUnet and Swin-Unet according to Recall based on their results (Equation 13–14).

$$F1 = DICE = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{13}$$

$$Recall = \frac{DICE * Precision}{2 * Precision - DICE} \tag{14}$$

Table 2. Dice, Precision and Recall for TransUnet and Swin-Unet based on caries segmentation on tooth X-ray images

	TransUnet	Swin-Unet
DICE	0.7096	0.7076
Precision	0.6837	0.7239
Recall	0.7375	0.6920

According to Table 2 TransUnet shows better Recall and DICE for this particular case.

So, when we observed models not only according to DICE and the number of parameters, TransUnet is preferable in some cases.

Conclusion

The shortage of specialized doctors is a pressing problem today. The healthcare system is forced to cope with an extremely high workload, which leads to irreversible consequences. Improving patient management can partially overcome this problem. Pre-diagnostic analysis can be used as a preliminary step before professional diagnosis. This way, patient queues can be managed before seeing a specialist. The use of artificial intelligence in pre-hospital diagnostics can greatly reduce the burden on the medical sector. One effective way to optimize pre-medical diagnostics is through the application of AI in the diagnosis of medical data, particularly in medical image processing. Convolutional Neural Networks have been widely used for this purpose. U-Net based CNN showed high performance and found its application for segmentation of different types of medical images. But their limitations have led to the use of more advanced models for computer vision, such as transformer-based models.

Transformer-based models allow for the viewing of the entire image at once, preserving spatial information that may be lost in the "sliding window" approach used by convolutional neural networks. This has led to an increase in the quality of segmentation, although transformers are more complex models and require consideration of the number of parameters.

In this review, we compared two widely cited Unet-based models: TransUnet and Swin-Unet. They have a fundamental difference in architecture. TransUnet is a hybrid model with a transformer encoder and convolutional decoder. Swin-Unet is a fully transformer model. This difference leads to an increase in the number of parameters for TransUnet. We examined the results provided by the original articles' authors and supplemented them by calculating the number of parameters according to the provided code. We demonstrate that the newer Swin-Unet showed better performance in terms of DICE, HU, and a range of parameters compared to TransUnet.

References

1. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., & Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541-551. https://doi.org/10.1162/neco.1989.1.4.541
2. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84-90. https://doi.org/10.1145/3065386
3. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009, June). ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848
4. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). IEEE.
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929
7. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18* (pp. 234-241). Springer International Publishing. https://doi.org/10.1007/978-3-319-24574-4_28
8. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., & Wang, M. (2023, February). Swin-unet: Unet-like pure transformer for medical image segmentation. In *Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III* (pp. 205-218). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-25066-8_9
9. Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., ... & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*. https://doi.org/10.48550/arXiv.2102.04306
10. Jia, Q., & Shu, H. (2022, July). Bitr-unet: a cnn-transformer combined network for mri brain tumor segmentation. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 7th International Workshop, BrainLes 2021, Held in Conjunction with MICCAI 2021, Virtual Event, September 27, 2021, Revised Selected Papers, Part II* (pp. 3-14). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-09002-8_1
11. Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.A., ... & Jodoin, P.M. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?. *IEEE transactions on medical imaging*, *37*(11), 2514-2525. https://doi.org/10.1109/TMI.2018.2837502
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 10012-10022).
13. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
14. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.https://doi.org/10.48550/arXiv.1409.1556
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117-2125).
16. *Multi-Atlas Labeling Beyond the Cranial Vault - Workshop and Challenge*. Synapse multi-organ computer tomography dataset. [Data set]. Synapse.org. https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn3193805type=ENTITY. https://doi.org/10.7303/SYN3193805
17. Chen et. al. (2021). *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*. Github. https://github.com/Beckschen/TransUNet

18. Hu et.al. (2022). *Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation*. Github. https://github.com/HuCaoFighting/Swin-Unet
19. Ying, S., Wang, B., Zhu, H., Liu, W., & Huang, F. (2022). Caries segmentation on tooth X-ray images with a deep network. *Journal of Dentistry*, *119*, 104076. https://doi.org/10.1016/j.jdent.2022.104076
20. Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*. https://doi.org/10.48550/arXiv.1803.02155