A study of attention information from transformer layers in hybrid medical image segmentation networks

Syed Nouman Hasany, Caroline Petitjean, Fabrice Meriaudeau

A study of attention information from transformer layers in hybrid medical image segmentation networks

Syed Nouman Hasany[a], Caroline Petitjean[a], and Fabrice Mériaudeau[b]

[a]LITIS UR 4108, Normandie Université, INSA de Rouen, UNIROUEN, UNIHAVRE, Rouen, France
[b]Institut de Chimie Moleculaire de l'Universite de Bourgogne, ICMUB UMR CNRS 6302, Universite Bourgogne, Dijon 21000, France

ABSTRACT

Transformer models have recently started gaining popularity in Computer Vision related tasks. Within Medical Image Segmentation, segmentation models such as TransUNet have incorporated transformer blocks alongside convolutional blocks while remaining faithful to the encoder-decoder based architecture popularized by U-Net. The rationale behind this is to supplement the local information obtained from convolutional kernels with the global information obtained from transformer blocks. The present work examines information flow with a focus on attention values within transformer blocks of three such segmentation models: (i) TransUNet, (ii) 2D CATS (Complementary CNN and Transformer Encoders for Segmentation), and (iii) 2D UNETR (UNEt TRansformers). For each of these models, an analysis of attention information reveals as to how many transformer blocks are necessary in order to effectively achieve a global receptive field. Based on this, compressed versions of these models are proposed, helping reduce the number of model parameters to around 40% of the original parameters for 2D CATS and around 25% for TransUNet and 2D UNETR. With the help of three different datasets (IBSR 18, EMIDEC, Synapse multi-organ), it is shown that in terms of the dice metric the performance of the compressed model does not drop by more than 5% compared to the original model.

Keywords: Segmentation Model Compression, Information Flow, Attention Maps, Hybrid Segmentation Models

1. INTRODUCTION

Following the success of AlexNet,[1] the last decade of Medical Image Analysis has been dominated by Convolutional Neural Networks (CNNs). Recently, however, an alternate approach towards image analysis has been proposed which utilizes transformer blocks.

Transformer blocks differ significantly from convolutional blocks in that they can theoretically learn global relationships within an image. Convolutional blocks, on the other hand, can only extract local information. This difference primarily stems from the fact that a convolutional block effectively has a limited field of view (i.e. receptive field) whereas for a conventional transformer block, the field of view is essentially the entire image. While originally proposed in the context of machine translation,[2] transformer blocks were introduced within computer vision via the Vision Transformer (ViT)[3] architecture in the context of image classification.

Unlike image classification where the ViT[3] relied solely on transformer blocks, image segmentation models incorporating transformer blocks have also tended to incorporate convolutional blocks. As far as medical image segmentation is concerned, the encoder-decoder based U-Net architecture[4] and its variants have been the relied upon workhorse. Segmentation models incorporating transformer blocks, too, have continued with this basic encoder-decoder U-shaped architecture and three such models are the subject of the present study:

- TransUNet[5]

Further author information: (Send correspondence to S.N.H)
S.N.H.: E-mail: syednoumanhasany1997@gmail.com,
F.M.: E-mail: fabrice.meriaudeau@u-bourgogne.fr

- CATS[6]
- UNETR[7]

U-Net[4] (and models following its general architecture) utilizes an encoder which consists of multiple convolutional stages, each of which involves convolutional blocks followed by a pooling operation to reduce the spatial dimensions. The rationale behind this approach is to consider the receptive field as it increases with each successive stage, and by the time the bottleneck block is reached, the model often achieves a global receptive field.

In comparison, transformer based models such as ViT[3] which utilize the self-attention mechanism can, at least theoretically, achieve an effectively global receptive field from the first block. However, since their is no guarantee that it will be achieved within the first block, visualizing attention information can serve as a guide as to how the effective receptive field changes as information flows through transformer blocks. We can utilize this attention information to identify the minimum number of blocks required in order to achieve an effectively global receptive field. Once identified, a compressed version of the transformer based segmentation model can be proposed which only retains the appropriate number of transformer blocks.

There are three main components of the present paper. The first is to utilize segmentation models incorporating transformer blocks based on the encoder-decoder architecture and train them on three medical imaging datasets:

- Synapse multi-organ dataset
- EMIDEC dataset
- IBSR 18 dataset

The second component involves working with model interpretability techniques in order to visualize information flow within such architectures with a focus on the self-attention mechanism. Lastly, the final component utilizes results from this investigation in order to compress the studied architectures, and compare the performance of the compressed versions to the originals.

2. BACKGROUND

Image classification was dominated by CNNs in the last decade based on the success of architectures such as AlexNet,[1] ResNet,[8] and EfficientNet.[9] Inspired from Natural Language Processing (NLP), however, a transformer based model having no convolutional blocks was proposed in 2020 - the Vision Transformer model (ViT).[3] It starts by dividing an image into patches. Embeddings extracted from each patch - similar to word token embeddings in NLP - are passed on to a series of transformer blocks before adding a multilayer perceptron to the final layer for a classification decision. The same paper also proposed a hybrid model in which instead of the transformer blocks directly operating on the original image, they are applied to a feature representation of the original image obtained from a CNN based backbone.

Similar to image classification, image segmentation was also dominated by CNNs. The encoder-decoder based U-Net[4] was a relatively popular model inspiring derivatives such as V-Net[10] and U-Net++.[11] Attention was also utilized in some derivatives such as the Attention U-Net,[12] Attention Unet++,[13] and Attention Gated Network.[14] The earliest incorporation of transformer blocks, however, was done in TransUNet in 2021.[5] The idea behind TransUNet is to replace the bottleneck convolutional layer of a U-Net (with a ResNet backbone as an encoder) with transformer blocks.

Following TransUNet, other segmentation models incorporating transformer blocks were also proposed such as the TransBTS[15] which expanded upon TransUNet to be directly applicable to 3D images, LeViT-UNet[16] which replaced the ViT transformer in TransUNet with LeViT, CoTr[17] in which the transformer blocks were applied not only to the bottleneck layer but to the remaining layers of the multi-scale feature map as well, Swin UNETR[18] in which the U-Net encoder was replaced by Swin transformer blocks, etc.

Many of the proposed models follow the encoder-decoder based U-Net architecture and incorporate transformer blocks on the encoder side. While TransUNet replaces the bottleneck layer with transformer blocks, other models such as CATS[6] introduce a transformer based parallel path, the information from which is fused with that coming from the convolutional path before getting passed on to the decoder. Another model is the UNETR[7] which gets rid of all convolutional blocks within the encoder - barring one - and replaces them with transformer blocks.

3. MATERIAL AND METHODS

3.1 Transformer

The transformer model was initially proposed in the context of machine translation.[2] The self-attention based transformer model can essentially be thought of as a representation learning mechanism for sequential data. With every successive transformer block, the representation of individual sequence units is modified taking all other sequence units into account.

An individual transformer block generally consists of two layers, a self-attention layer and a multilayer perceptron. The input to the transformer block is an input sequence of length N, with each unit of the sequence being represented by an embedding of size D. In order to inject positional information into the model, a positional embedding is added to each input embedding. These positional embeddings can either be pre-determined or learnt during the training process. For each input embedding, three vectors of size d_k representing the "key", "query", and the "value" are obtained via a simple matrix multiplication involving the input embedding and the key, query, and value matrices respectively. These matrices are learnt during the training process. For each input unit, it is determined as to which input units (including itself) should contribute towards its next representation. This is achieved by taking a dot product between the query vector of the concerned input unit and the key vector of all units in the input sequence. A softmax is then applied to a scaled version of this dot product representing the importance of each unit towards the unit in question. The representation is then formed using a linear combination of value vectors such that the results of the softmax form the coefficients of the linear combination. This completes the self-attention mechanism. Self-attention can be applied in a single step for the entire sequence as expressed in the following equation:

$$Attention(Q, K, V) = softmax(\frac{Q \cdot K^T}{\sqrt{d_k}}) \cdot V \tag{1}$$

Where Q, K, V represent the query, key, and value representations of the entire sequence, each of shape $N \times d_k$.

In order to allow for multiple useful representations, instead of using a single self-attention mechanism, a multi-head self-attention mechanism is utilized. Put simply, self-attention is repeated multiple times for each unit, and the eventual representations are concatenated to obtain a final representation. Following self-attention, the representations are passed through a multilayer perceptron whose weights are shared between all units of the sequence.

In addition to the two layers, each transformer block also makes use of layer normalization and residual connections. Both self-attention, and multilayer perceptron are preceded with layer normalization, and succeeded with residual connections. The entire workflow of a self-attention based transformer block can, thus, be expressed in the following set of equations:

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

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

Where MSA refers to multi-head self-attention, MLP refers to multilayer perceptron, z_{l-1} represents the representations from the preceding block, and z_l represents the representations from the current block.

Following the success of transformers in the field of Natural Language Processing,[3] successfully applied them in the field of Computer Vision, particularly image classification. Since images, unlike textual data, are not inherently sequential, the authors utilized two techniques via which image data can be made compatible with transformers:

- Raw image patches
- Hybrid architectures

3.1.1 Raw Image Patches

The authors propose splitting the image into non-overlapping image patches. Each patch is then flattened, and passed through an embedding layer. The sequence of these input embeddings is what forms the input to the transformer model. Conventionally, for an image of size 224×224, the patch size is taken to be 16×16 leading to 14×14 patches which form 196 input tokens.

In addition to this, since the vision transformer model is also required to make a classification decision, an extra token is added in the beginning of the sequence - referred to as the "CLS" token. With each transformer block the representation of the input embeddings keep getting modified, and from the last transformer block, the representation corresponding to the "CLS" position is passed through a simple multilayer perceptron followed by a softmax to get a classification decision.

3.1.2 Hybrid Architectures

An alternative to raw image patches is to avoid applying the transformer model directly to the input image. Instead, the authors propose passing the input image through a Convolutional Neural Network, and the transformer model can then be applied to the output of an intermediate convolutional layer. The authors utilize the ResNet family of architectures as their backbone Convolutional Network. The rationale behind this method can be two-fold. Firstly, the spatial dimensions of the intermediate feature maps would have decreased considerably by then, allowing the transformer to be applied such that each pixel position be considered as an individual input. Secondly, this allows for the convolutional network to provide the transformer with an already rich feature representation as its input.

3.2 Segmentation Models incorporating Transformer Blocks based on the Encoder-Decoder Architecture

ViT's success has led to the application of transformer models into domains other than image classification, a prime example being that of Image Segmentation. Many transformer based image segmentation models closely follow the encoder-decoder based architecture supplemented with skip connections. Presently, the focus will be on three such architectures each of which modify the encoder portion of a conventional U-Net in a unique manner:

- TransUNet - replaces the bottleneck convolutions of the encoder with transformer blocks
- CATS - runs the image through both transformer blocks and convolutions separately and combines the information at each encoder step
- UNETR - replaces the convolutional part of the encoder entirely with transformer blocks barring one convolutional layer

3.2.1 TransUNet

TransUNet makes use of a hybrid vision transformer.[3] The encoder consists of feature maps obtained from a ResNet based backbone, the last of which is reshaped and fed to 12 transformer blocks. The detailed architecture can be seen in Figure. 1. It is worth noting that a major difference between this transformer model and the transformer models used in image classification is the lack of an extra input token in the beginning as we are not interested in a classification decision, but are only interested in utilizing the representations obtained from the transformer.

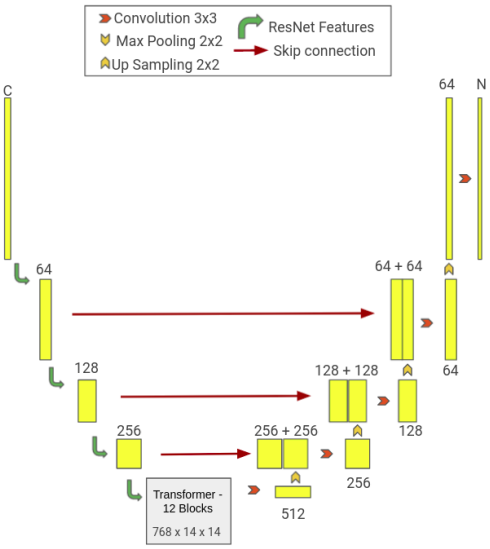

Figure 1: TransUNet architecture - the encoder side - a ResNet-50 based backbone - processes the image gradually decreasing its spatial dimensions. Once the spatial dimensions reach 14×14, the features are processed by a transformer, the output of which is passed on to the decoder which performs the process of upsampling and concatenating skip connections from the corresponding encoder output. The numbers in the figure represent the channels from each stage of the process

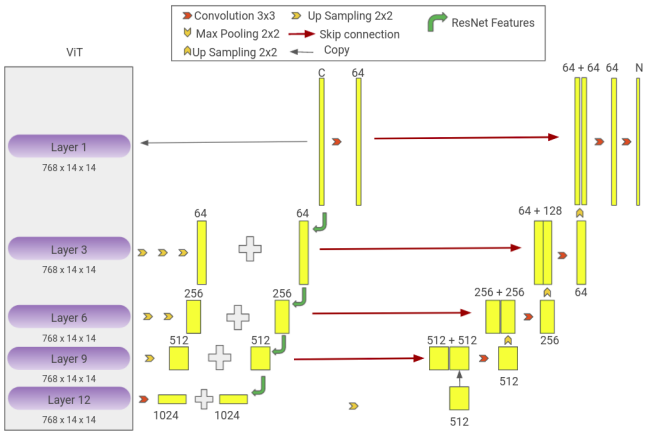

Figure 2: 2D CATS - the encoder side is composed of two parallel pathways. One is composed of a convolutional backbone - a ResNet-50 in this figure - which processes the image by gradually decreasing its spatial dimensions. The other one is composed of a pre-trained ViT which works on 196 (14×14) non-overlapping patches (size 16×16) from the original image and the output from each of its blocks can be reshaped to a spatial dimension of 14×14. The numbers in the figure represent the channels from each stage of the process

3.2.2 CATS

Introduced in 2022, CATS[6] retains the original convolution based U-Net encoder. However, it adds an additional path in which the original image is passed through a transformer model. The information from the convolutional pathway and the transformer pathway is then fused using simple addition before flowing on to the decoder. While the original CATS paper proposed a network for 3D images, the present work presents a slightly modified version which is supposed to work for 2D images. In addition, unlike the original CATS, the present network utilizes a pre-trained ViT model. The detailed architecture can be seen in Figure. 2.

3.2.3 UNETR

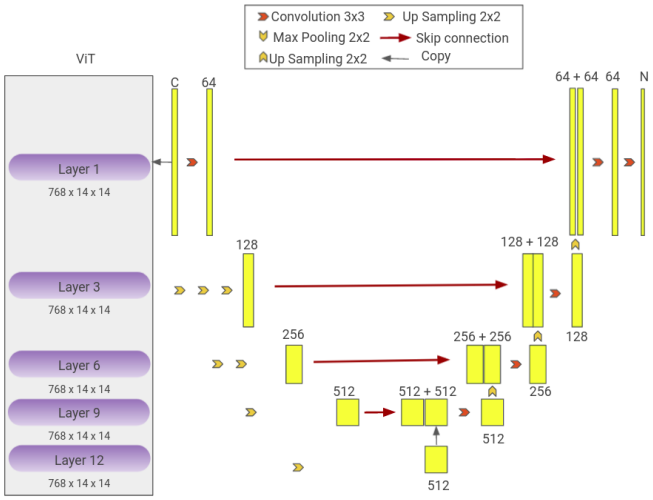

Figure 3: 2D UNETR - the encoder side is composed of a pre-trained ViT which works on 196 (14×14) non-overlapping patches (size 16×16) from the original image and the output from each of its blocks can be reshaped to a spatial dimension of 14×14. There is also an independent branch connecting the image directly to the final decoder layer. The numbers in the figure represent the channels from each stage of the process

Introduced in 2021, UNETR[7] consists of an encoder which is almost entirely based on transformer blocks. Similar to CATS, in the present work the original architecture has been modified such that it works for 2D images instead of 3D, and it utilizes a pre-trained ViT model instead of randomly initializing the transformer blocks. The detailed architecture can be seen in Figure. 3.

3.3 Information Flow

For transformers, attention values provide us with the clue as to where attention was being paid in a particular transformer block. The present work considers two approaches for this task. A relatively simple approach in this regard is to observe the raw attention values from each block.

3.3.1 Raw Attention Values

Transformer models afford us the possibility to not just observe feature maps, but to also observe how much each patch in any block was attending to the other patches (including itself). This can be achieved by visualizing the raw attention values obtained from the dot-product of key and query matrices.

3.4 Datasets

Experiments were performed using three different medical image segmentation datasets. The first one is the IBSR 18 dataset (http://www.nitrc.org/projects/ibsr). It contains the T1-weighted brain MRI images for 15 patients. 10 patients were used for training and 5 for validation leading to 1280 2D slices for training and 640 for validation. The labels are Cerebrospinal Fluid (CSF), Gray Matter (GM), and White Matter (WM).

The second dataset is the Synapse multi-organ dataset (https://www.synapse.org/Synapse:syn3193805/wiki/217789). 30 cases of abdominal CT scans are provided, 18 of which are used for training and 12 for validation leading to 2211 2D slices for training and 1568 for validation. The labels utilized are Aorta, Gallbladder, Spleen, Left Kidney, Right Kidney, Liver, Pancreas, and Stomach.

The third dataset is the EMIDEC Challenge dataset[19] (http://emidec.com/dataset). 100 cases of delayed-enhancement cardiac MRI are provided, 80 of which are used for training and 20 for validation leading to 558 2D slices for training and 180 for validation. The labels are Myocardium, Infarction, and NoReflow.

3.5 Pre-processing

The datasets were normalized with each image volume ending up with zero mean and unit standard deviation. For the Synapse multi-organ dataset, dataset is obtained from the TransUNet[5] authors in which, prior to normalizing, the images were clipped within a range of -125 and 275. For the EMIDEC dataset, center cropping was performed with a size of 96×96. Eventually, each slice from all images is resized to 224×224.

3.6 Training Configuration

All models were trained with an AdamW optimizer[20] with a learning rate of $1e-04$ without any weight decay. For the IBSR 18 and Synapse multi-organ datasets, 80 epochs were utilized whereas for the EMIDEC dataset, 100 training epochs were utilized. In each case, the batch size was 8. The loss function in each case was an average of dice loss and cross-entropy loss.

3.7 Analyzing Information Flow

3.7.1 Attention Information

In order to analyze attention flow for each model, raw attention maps were visualized for all transformer blocks. In general, at each step there is a sequence of length 196, and each unit of that sequence pays attention to each other unit implying 196 attention values per unit. For every unit, an attention map of size 14×14 can be obtained, and since there are 196 units, each of these 14×14 attentions maps can be displayed on a 14×14 grid. In each case raw attention maps from the first, second, third, and sixth transformer block are visualized.

An analysis of raw attention maps for TransUNet (Figure. 4) reveals that, generally, starting from the first transformer block, patches start paying attention to the remaining patches irrespective of their spatial proximity to those patches. For example, for a shape of 14×14, position $(1, 1)$ can potentially pay attention to a patch at position $(14, 14)$. This trend continues for the remaining transformer blocks as well.

Analyzing the raw attention maps from the 2D UNETR (Figure. 5) reveals an interesting pattern. In the first block each patch mostly seems to be paying attention only to itself. For the second block, attention maps reveal that each patch is mostly paying attention to patches which are within close proximity. This behaviour closely resembles that of a convolutional kernel passing over an image. From the fourth block onward, however, the attention values reveal a more global trend. 2D CATS displays a similar trend as can be referred to from the codebase: *link_to_add*

3.8 Model Compression

An analysis of attention maps indicate that the architectures under consideration might lend themselves to compression. In a conventional U-Net, an encoder contains multiple convolutional blocks primarily because as one goes from one block to the next, the receptive field of the model increases and, hence, more context can be incorporated. If one contrasts it with the attention information obtained from the 2D UNETR and 2D CAT models, it can be seen that starting from the third transformer block, attention maps already seem to indicate a global receptive field. Taking this information into account, compressed versions of both 2D CATS and 2D UNETR are proposed as can be seen in Figures 6 and 7 respectively.

For the TransUNet, attention maps seem to indicate that spatial proximity plays no role even in the first and second transformer blocks. Hence, two compressed versions are proposed. In the first version, instead of having twelve transformer blocks, the model has only three. In a second version the model is further compressed to have only one transformer block.

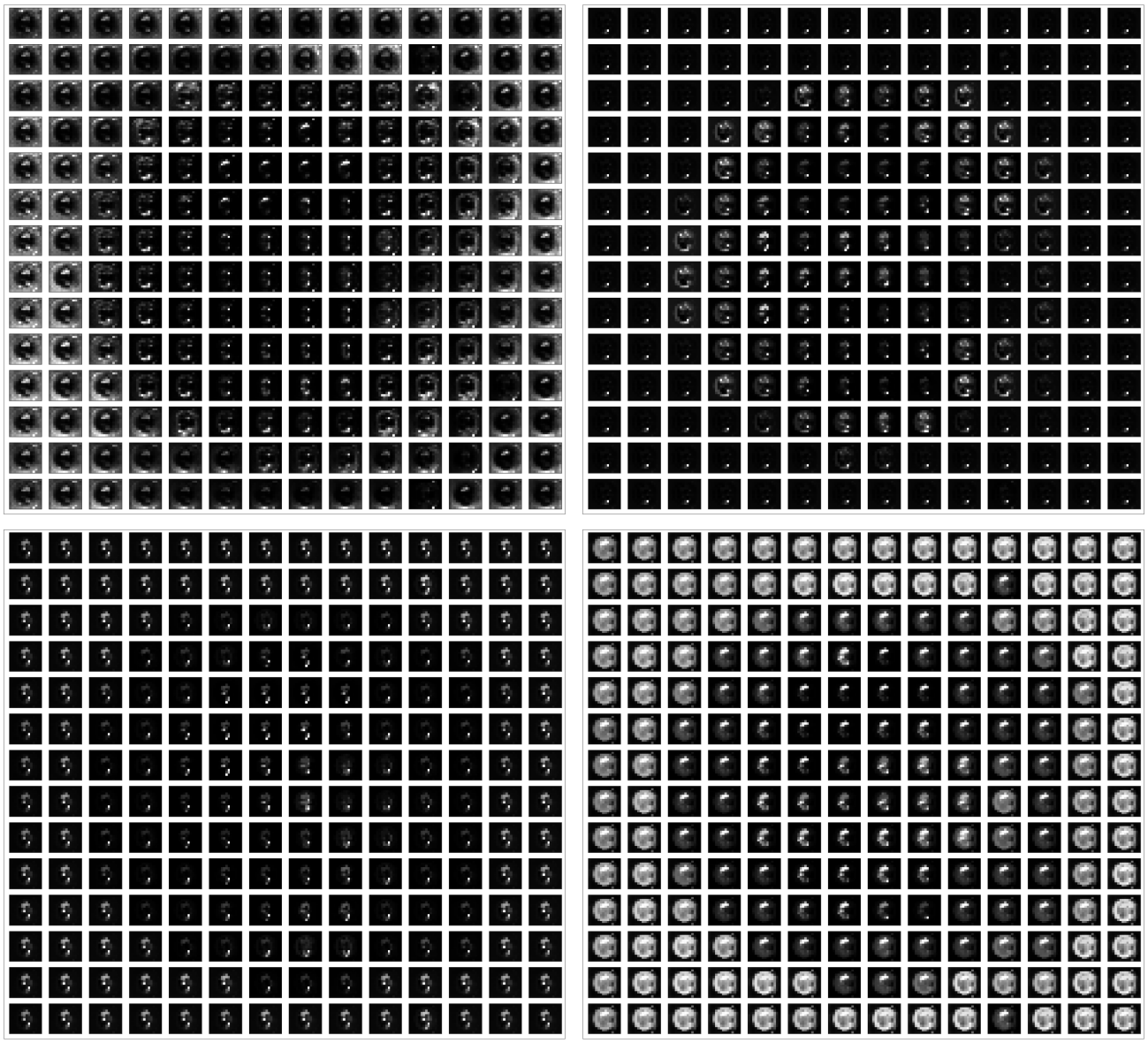

Figure 4: TransUNet - raw attention maps for a sample image from Synapse multi-organ dataset. Image reads from left to right and top to bottom. Maps for the first (top-left), second (top-right), third (bottom-left), and sixth (bottom-right) transformer blocks are plotted

Model	Transformer Blocks	Dataset Size	CSF	Gray Matter	White Matter	Average Dice	% Dice Change (Compressed vs Uncompressed)	% Parameter Saving (Compressed vs Uncompressed)
TransUNet	12	1200	0.816	0.894	0.870	0.860	-	-
TransUNet	3	1200	0.809	0.897	0.874	0.860	0	60.59
TransUNet	1	1200	0.828	0.896	0.873	0.865	0.58	74.06
2D CATS	12	1200	0.820	0.893	0.873	0.862	-	-
2D CATS	3	1200	0.812	0.902	0.877	0.864	0.19	61.15
2D UNETR	12	1200	0.815	0.901	0.883	0.866	-	-
2D UNETR	3	1200	0.814	0.902	0.877	0.864	-0.21	75.12

Table 1: Dice metrics for IBSR 18 dataset. For the 2D CATS and 2D CATS - Compressed models, the backbone utilized was ResNet-50

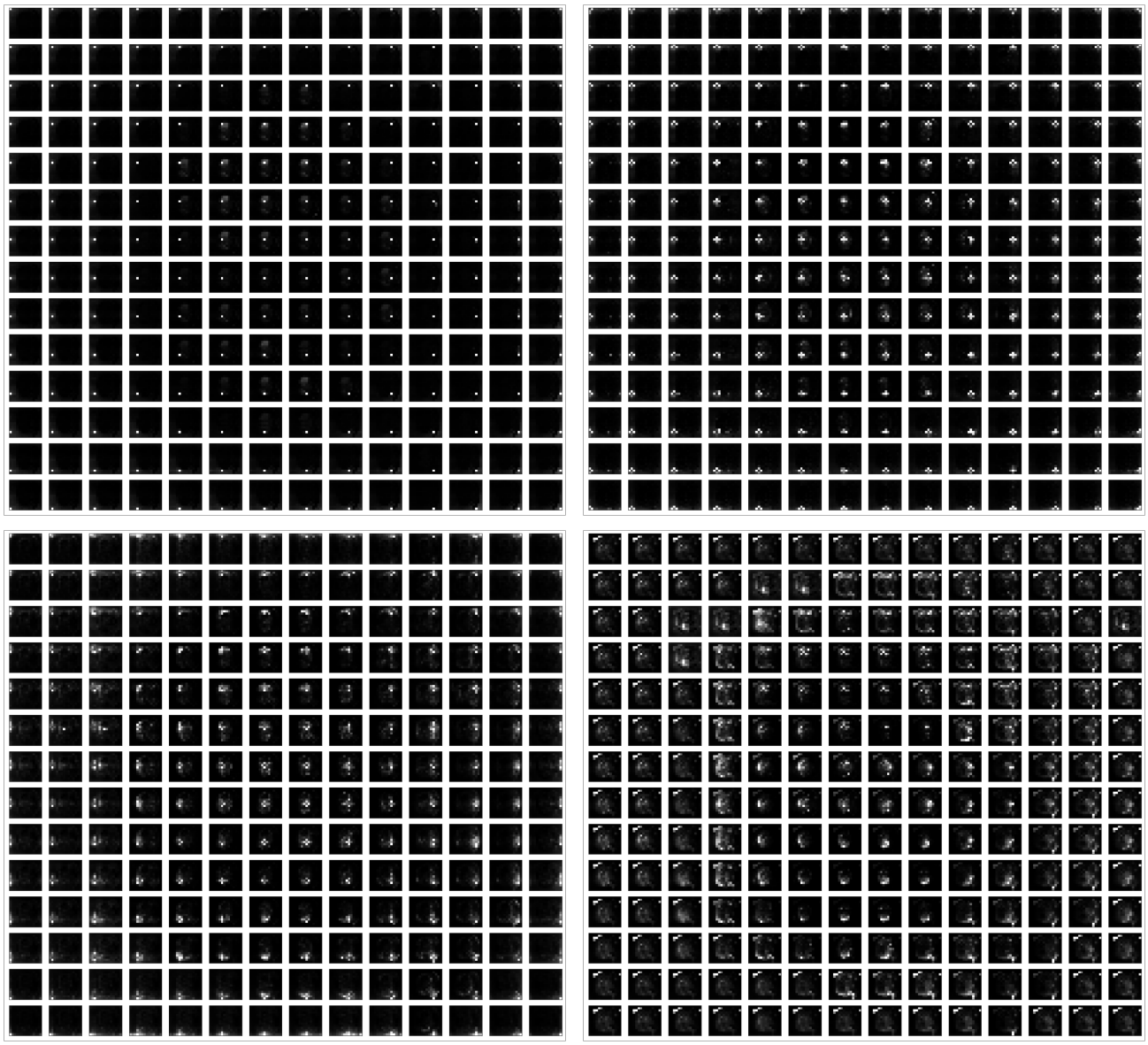

Figure 5: 2D UNETR - raw attention maps for a sample image from Synapse multi-organ dataset. Image reads from left to right and top to bottom. Maps for the first (top-left), second (top-right), third (bottom-left), and sixth (bottom-right) transformer blocks are plotted

Model	Transformer Blocks	Dataset Size	Myocardium	Infarction	NoReflow	Average Dice	% Dice Change (Compressed vs Uncompressed)	% Parameter Saving (Compressed vs Uncompressed)
TransUNet	12	558	0.846	0.654	0.774	0.758	-	-
TransUNet	3	558	0.853	0.676	0.749	0.760	0.25	60.59
TransUNet	1	558	0.859	0.680	0.767	0.769	1.46	74.05
2D CATS	12	558	0.855	0.569	0.674	0.699	-	-
2D CATS	3	558	0.846	0.661	0.652	0.720	2.90	65.42
2D UNETR	12	558	0.841	0.593	0.723	0.719	-	-
2D UNETR	3	558	0.830	0.617	0.663	0.703	-2.16	75.12

Table 2: Dice metrics for EMIDEC dataset. For the 2D CATS and 2D CATS - Compressed models, the backbone utilized was ResNet-34

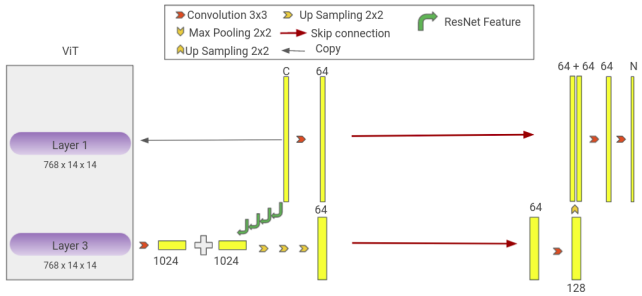

Figure 6: 2D CATS - Compressed - only the first three transformer blocks from the original 2D CATS are retained. The numbers in the figure represent the channels from each stage of the process

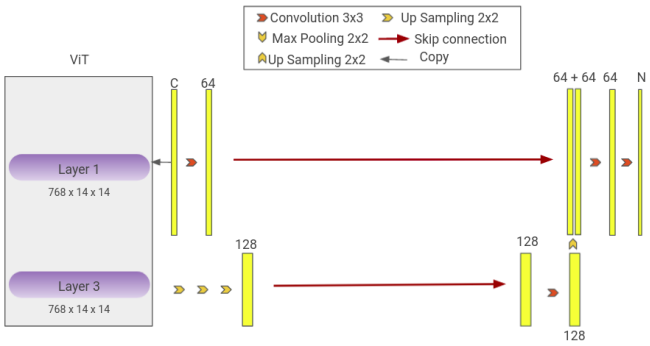

Figure 7: 2D UNETR - Compressed - only the first three transformer blocks from the original 2D UNETR are retained. The numbers in the figure represent the channels from each stage of the process

Model	Transformer Blocks	Dataset Size	Aorta	Gallbladder	Kidney (L)	Kidney (R)	Liver	Pancreas
TransUNet	12	2212	0.873	0.662	0.869	0.826	0.944	0.612
TransUNet	3	2212	0.878	0.667	0.874	0.842	0.948	0.665
TransUNet	1	2212	0.881	0.643	0.859	0.831	0.947	0.618
2D CATS	12	2212	0.848	0.664	0.857	0.818	0.935	0.626
2D CATS	3	2212	0.862	0.671	0.855	0.803	0.939	0.567
2D UNETR	12	2212	0.853	0.693	0.861	0.773	0.947	0.603
2D UNETR	3	2212	0.853	0.656	0.832	0.760	0.950	0.568

Model	Transformer Blocks	Dataset Size	Spleen	Stomach	Average Dice	% Dice Change (Compressed vs Uncompressed)	% Parameter Saving (Compressed vs Uncompressed)
TransUNet	12	2212	0.912	0.843	0.818	-	-
TransUNet	3	2212	0.891	0.826	0.824	0.73	60.59
TransUNet	1	2212	0.916	0.836	0.816	-0.15	74.05
2D CATS	12	2212	0.893	0.838	0.810	-	-
2D CATS	3	2212	0.879	0.732	0.788	-2.65	63.12
2D UNETR	12	2212	0.903	0.768	0.800	-	-
2D UNETR	3	2212	0.888	0.783	0.786	-1.72	75.12

Table 3: Dice metrics for Synapse-multiorgan dataset. For the 2D CATS and 2D CATS - Compressed models, the backbone utilized was DenseNet-121. The table has been divided in two parts in order to ease visibility

4. RESULTS

The dice metrics for the uncompressed and the compressed models can be seen in Table. 1 for IBSR 18 dataset, Table. 2 for EMIDEC dataset, and Table. 3 for Synapse multi-organ dataset.

In addition to the dice metrics, the tables also indicate the percentage change in model performance going from an uncompressed model to a compressed one:

$$\% \; Change = 100 \times \frac{PC - PuC}{PuC} \tag{4}$$

Where PC is the performance of the compressed model and PuC is the performance of the uncompressed model.

Lastly, the table also indicates the percentage saving in model parameters going from an uncompressed model to a compressed one:

$$\% \; Change = 100 \times \frac{\#uC - \#C}{\#uC} \tag{5}$$

Where $\#C$ are the number of parameters in the compressed model and $\#uC$ are the number of parameters in the uncompressed model.

5. DISCUSSION AND CONCLUSIONS

An analysis of raw attention maps revealed that it is not necessary to have all 12 transformer blocks in order to achieve a global receptive field. Based on this analysis, model compression was performed such that all blocks after the first such block which achieves a global receptive field were discarded. The compressed versions of all three original models (TransUNet, 2D CATS, 2D UNETR) have less than 50% of the original parameters. The compressed version of 2D UNETR and the single block TransUNet have, in fact, less than one-third of the original parameters. It can thus be argued that attention information from transformer blocks is helpful not only towards analyzing information flow, but it can also influence architectural decisions leading to model compression without seriously sacrificing model performance.

6. ACKNOWLEDGMENTS

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR), under grant Project-ANR-21-CE23-0013 (project MediSEG).

REFERENCES

[1] Krizhevsky, A., Sutskever, I., and Hinton, G. E., "Imagenet classification with deep convolutional neural networks," *Commun. ACM* , 84–90 (2017).

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, N. A., Kaiser, L., and Polosukhin, I., "Attention is all you need," *Advances in Neural Information Processing Systems 30 (NIPS 2017)* , 5998–6008 (2017).

[3] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N., "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR* (2021).

[4] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science* , 234–241 (2015).

[5] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, L. A., and Zhou, Y., "Transunet: Transformers make strong encoders for medical image segmentation," (2021).

[6] Li, H., Hu, D., Liu, H., Wang, J., and Oguz, I., "Cats: Complementary cnn and transformer encoders for segmentation," in [*2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*], 1–5 (2022).

[7] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, A. B., Roth, R. H., and Xu, D., "Unetr - transformers for 3d medical image segmentation," *WACV* , 1748–1758 (2022).

[8] He, K., Zhang, X., Ren, S., and Sun, J., "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* , 770–778 (2016).

[9] Tan, M. and Le, V. Q., "Efficientnet: Rethinking model scaling for convolutional neural networks," *International Conference on Machine Learning* , 6105–6114 (2019).

[10] Milletari, F., Navab, N., and Ahmadi, S.-A., "V-net: Fully convolutional neural networks for volumetric medical image segmentation," *Proceedings of 2016 Fourth International Conference on 3D Vision (3DV)* , 565–571 (2016).

[11] Zhou, Z., Siddiquee, M. R. M., Tajbakhsh, N., and Liang, J., "Unet plus plus : A nested u-net architecture for medical image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, DLMIA 2018* , 3–11 (2018).

[12] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, C. H. M., Heinrich, P. M., Misawa, K., Mori, K., McDonagh, G. S., Hammerla, Y. N., Kainz, B., Glocker, B., and Rueckert, D., "Attention u-net: Learning where to look for the pancreas," *arXiv: Computer Vision and Pattern Recognition* (2018).

[13] Li, C., Tan, Y., Chen, W., Luo, X., Gao, Y., Jia, X., and Wang, Z., "Attention unet++: A nested attention-aware u-net for liver ct image segmentation," *2020 IEEE International Conference on Image Processing (ICIP)* , 345–349 (2020).

[14] Schlemper, J., Oktay, O., Schaap, M., Heinrich, P. M., Kainz, B., Glocker, B., and Rueckert, D., "Attention gated networks: Learning to leverage salient regions in medical images," *Medical Image Analysis* , 197–207 (2019).

[15] Wang, W., Chen, C., Ding, M., Yu, H., Zha, S., and Li, J., "Transbts: Multimodal brain tumor segmentation using transformer," *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, PT I* , 109–119 (2021).

[16] Xu, G., Wu, X., Zhang, X., and He, X., "Levit-unet: Make faster encoders with transformer for medical image segmentation," (2021).

[17] Xie, Y., Zhang, J., Shen, C., and Xia, Y., "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021, PT III* , 171–180 (2021).

[18] Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., and Xu, D., "Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images," (2022).

[19] Lalande, A., Chen, Z., Decourselle, T., Qayyum, A., Pommier, T., Lorgis, L., Rosa, d. l. E., Cochet, A., Cottin, Y., Ginhac, D., Salomon, M., Couturier, R., and Meriaudeau, F., "Emidec: A database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri," *international conference on data technologies and applications* (2020).

[20] Loshchilov, I. and Hutter, F., "Decoupled weight decay regularization," *ICLR* (2019).