Analyzing Components of a Transformer under Different Data Scales in 3D Prostate CT Segmentation

Yicong Tan[a], Prerak Mody[b], Viktor van der Valk[b], Marius Staring[b,c], and Jan van Gemert[a]

[a]Pattern Recognition Lab, TU Delft, Delft, The Netherlands
[b]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands
[c]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands

ABSTRACT

Literature on medical imaging segmentation claims that hybrid UNet models containing both Transformer and convolutional blocks perform better than purely convolutional UNet models. This recently touted success of Transformers warrants an investigation into which of its components contribute to its performance. Also, previous work has a limitation of analysis only at fixed data scales as well as unfair comparisons with others models where parameter counts are not equivalent. This work investigates the performance of the window-Based Transformer for prostate CT Organ-at-Risk (OAR) segmentation at different data scales in context of replacing its various components. To compare with literature, the first experiment replaces the window-based Transformer block with convolution. Results show that the convolution prevails as the data scale increases. In the second experiment, to reduce complexity, the self-attention mechanism is replaced with an equivalent albeit simpler spatial mixing operation i.e. max-pooling. We observe improved performance for max-pooling in smaller data scales, indicating that the window-based Transformer may not be the best choice in both small and larger data scales. Finally, since convolution has an inherent local inductive bias of positional information, we conduct a third experiment to imbibe such a property to the Transformer by exploring two kinds of positional encodings. The results show that there are insignificant improvements after adding positional encoding, indicating the Transformers deficiency in capturing positional information given our data scales. We hope that our approach can serve as a framework for others evaluating the utility of Transformers for their tasks. Code is available via https://github.com/prerakmody/window-transformer-prostate-segmentation.

Keywords: Radiotherapy, Segmentation, Window-Based Transformer, Convolution, Pooling, Positional Encoding

1. INTRODUCTION

Cancer treatment via radiotherapy requires the segmentation of tumors and organs-at-risk (OAR) on diagnostic scans like CT. Convolution-based UNet architectures[1] have dominated this task for years. However, the recent success of Transformers in the vision domain has raised the question whether they can replace convolutions as the primary image processing operation in deep learning.[2–4] Specifically in the 3D medical imaging domain, the window-based Transformer[2] has been used since the vanilla Transformer has a computational complexity quadratic to the image size. However, owing to the scarcity of medical data, one of the deficiencies of such works is that their data sets are small, leading to potential overfitting.[2,3] Also, it is well known that Transformers need large datasets to perform well. Our work remedies this by analysing transformers in a UNet-based architecture at six different data scales. Inspired by ConvNeXt[5] we analyse different components of a transformer and replace them with more traditional deep learning operations like convolutions and pooling while at the same time ensuring equivalent parameter count and similar neural architectures. The results indicate that window-based Transformers perform worse than the comparison model in all our data scales for 3D prostate CT segmentation. Perhaps, Transformers need to evolve further to replace convolutions in the medical segmentation domain.

Further author information: (Send correspondence to P.P.M.)
P.P.M.: E-mail: P.P.mody@lumc.nl

2. METHOD

2.1 Data

We use prostate CT data containing annotations for four organs: bladder, prostate, rectum, and seminal vesicles. The data is collected from three institutes c.f. Haukeland Medical Center of Norway (HMC), Leiden University Medical Center of Netherlands (LUMC) and Erasmus Medical Center of Netherlands (EMC) containing 179, 475 and 56 CT scans respectively. EMC is used as the test data set while HMC and LUMC are used as the training datasets. Due to differences in clinical protocols for CT scan acquisition, the EMC dataset has larger volumes of the prostate and bladder which makes for a challenging test dataset.

2.2 Network Architecture

The Window-based Transformer network in evaluation is nnFormer,[2] which employs Swin (shifted-window)-Transformer blocks in the encoder and decoder of a UNet architecture. Note, that the first two layers of this architecture are convolution-based patch-embedding layers to extract low-level feature maps.

2.2.1 Method 1: Replacing Swin-Transformer block with convolution block

Literature on medical image segmentation has shown superior performance of window-based Transformers over convolutions.[2,3] We test this notion by replacing the window-based Transformer blocks with a sequence of two convolutions. We also ensure their parameter counts are equivalent and proceed to compare these models across multiple data scales. It is hypothesized that the Transformer will perform poorly in low-data regimes, since its attention mechanism is incapable of understanding relative position information of voxels, a quality important for precise tasks like segmentation and inherent to convolutions. Conversely, the lack of an inherent prior for imaging data, may allow Transformers to learn complex dependencies in the large-data regime and hence boost performance.

2.2.2 Method 2: Replacing the Self-Attention with Pooling

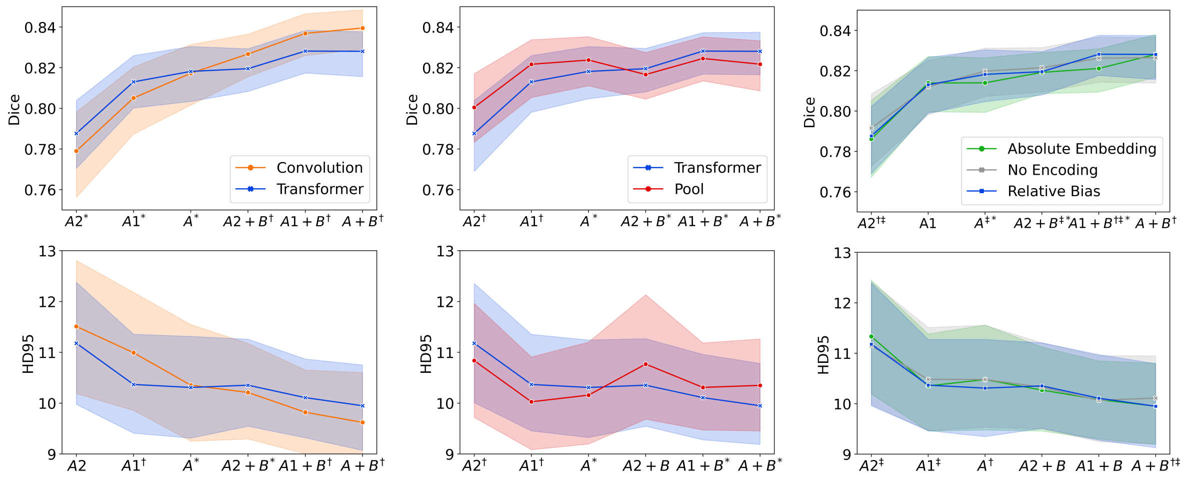
In the spirit of further analyzing components of the Transformer block and inspired by[6] to reduce computational complexity, we replace the attention mechanism with a much simpler spatial feature mixing operation, i.e pooling. Replacing the complex attention mechanism with a simpler pooling operation, may also reduce the chance of overfitting in low-data regimes. We hypothesize that max pooling will outperform self-attention in small data scales while self-attention will prevail gradually with increased data scale. This is because the complex nature of the attention mechanism when compared to max pooling might allow it to model spatial features provided additional data.

2.2.3 Method 3: Evaluating Positional Encoding

Under the assumption that failures of window-based Transformers might be due to its inability to model positional dependencies, we explore two different positional encoding methods and compare them with a model without any positional encodings. The first is absolute positional embedding that is added after the convolutional patch-embedding; the second is the relative positional bias that is added when computing the attention matrix in each Swin-Transformer block. Our base Transformer model uses relative positional bias which we expect to perform better as per work done in literature.[2]

3. EXPERIMENTS AND RESULTS

This work uses two datasets for training i.e. HMC (or clinic A) and LUMC (or clinic B). The HMC dataset is split into two parts for 2-fold cross validation and also creating smaller data scales. They contain 94 and 85 CT scans respectively and are henceforth referred to as A1 and A2. We make 6 combinations of these datasets c.f. A1, A2, A, A1+B, A2+B, A+B to create multiple data scales. Note, that the data from clinic B is not used for pretraining, but rather as additional scans during training. Three experiments are conducted to compare the window-based Transformer to its counterparts on two geometric metrics i.e. Dice and 95th percentile Hausdorff Distance (HD95) averaged over all scans of the test dataset. The models were trained using a combination of Dice and cross-entropy loss for 500 epochs. The window-based Transformer and convolution contain 158.49M and 155.85M parameters respectively. The CT scans were first resampled to the median spacing of each dataset and then randomly sampled patches of size (128,128,64) were input to the network. Models were trained with Pytorch 11.3 on a single Nvidia RTX6000 (24GB memory).



(a) Method 1 (b) Method 2 (c) Method 3

Figure 1. Line plots showing the mean and 95th percentile confidence interval of Dice and HD95. The x-axis denotes the various training data scales with clinics A and B while A2, A1 denote the first and second part of clinic A. For (a) and (b), $*$ denotes a statistical difference in at least one organ and † denotes a statistical difference on average and for at least one organ. In (c), †, ‡ and $*$ denote the statistical difference between the relative bias and no encoding, absolute embedding and no encoding, and the two positional encoding respectively on the average of all organs.

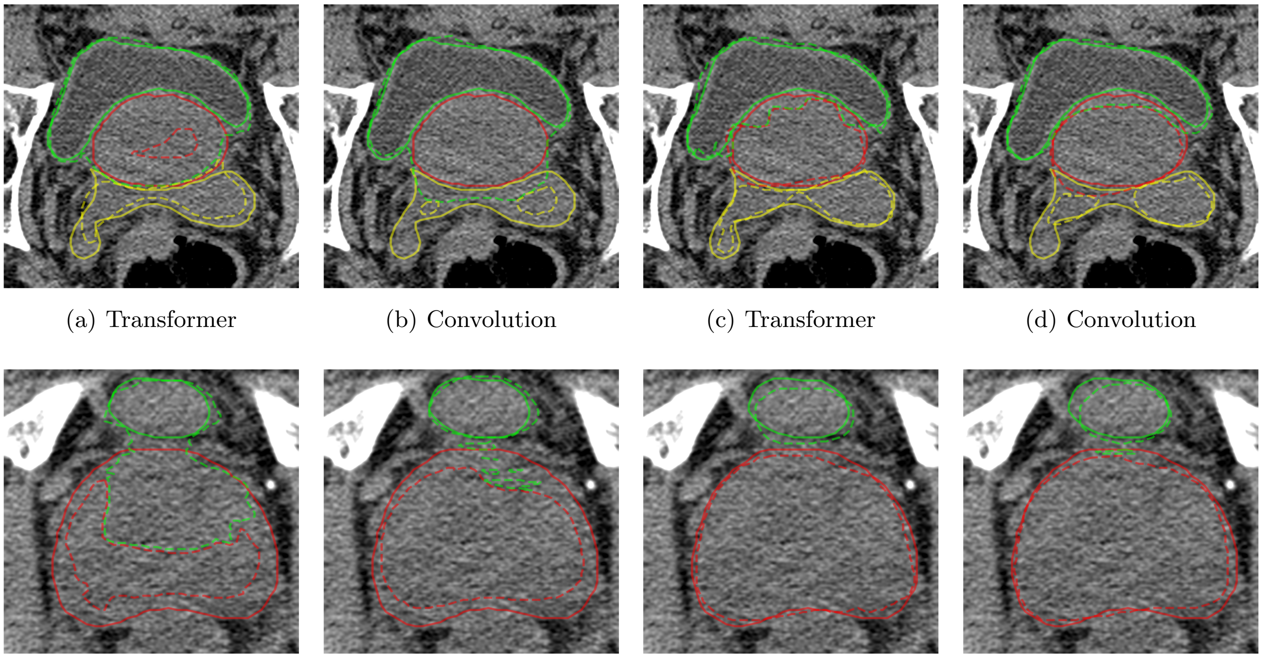
Surprisingly, Figure 1 (a) and Figure 2 (a)(b)(c)(d) show that the Transformer performs better at lower data scales and the convolution gradually surpasses it with increase of data. The convolution performs poorly in the lower-data regime since the lack of data coupled with its locality bias may not allow it to learn sufficient global shape-based information, but only local textural information in the neighbourhood of a voxel (*seminal vesicle in Figure 2 (b)*). A lower supervision loss during training and higher performance in cross-validation experiments on clinic A also are also indicators of the overfitting nature of convolutions in our smaller data regimes. The higher performance of convolutions in our larger data scales may imply that the Transformer needs more data to learn the dependencies within the data (*jagged nature of bladder in Figure 2 (c)*).

In line with our expectations, Figure 1 (b) and Figure 2 (e)(f)(g)(h) show that max-pooling outperforms in small data scales compared to the window-based self-attention mechanism and the latter surpasses with the increase of data scale. Thus, both convolution and window-based Transformer fail to be well-trained under small data scales. These results indicate that the simplicity of pooling may be essential to high performance in small data regimes.

Figure 1 (c) shows that in spite of statistical differences, the gaps in performance across the different positional encodings is not large. The lack of a large difference between the models with some form of positional encoding and those without, indicates that the current data scales are either insufficient to train the positional encodings well or that a better positional encoding design is needed for medical segmentation. Contrary to Transformers, both convolutions and pooling have some form of inductive bias (i.e. locality and neighbourhood structure). This could be one reason that the window-based Transformer is not the best choice in both our smaller and larger data scales.

4. DISCUSSION AND CONCLUSION

This study evaluates different components of the window-based Transformer to understand their role in its performance. Unlike previous work, we maintain a constant parameter count across models and also analyse the effect of the components under different data scales. Our results show that purely convolutional UNet models perform slightly better (*or at least as good as*) hybrid UNet models containing window-based Transformers



(a) Transformer (b) Convolution (c) Transformer (d) Convolution

(e) Transformer (f) Pooling (g) Transformer (h) Pooling

Figure 2. CT scans showing the prediction (dotted line) and ground truth (solid line) for the prostate (red), bladder (green) and seminal vesicle (yellow). (a),(b),(e),(f) and (c),(d),(g),(h) show results when trained on the smallest and largest data scale respectively.

and convolutions. This result suggests that under fair experimental settings, the claimed superiority of the Transformer does not hold for prostate CT OAR segmentation. which we believe might be due to the the lack of understanding of positional information of voxels in our data scales. The results of the pooling operation suggest that it may always be better to choose simpler operations in low data regimes. The comparable performance of models with and without positional encodings further supports our first claim. Thus, we conclude that for our dataset the window-based Transformer is not the best choice in both small and larger data scales. Note here, that our largest data scale may not be sufficient for Transformers which are well-known to be data hungry. Future work could use our approach to understanding Transformers by either segmenting other organs or using different medical imaging modalities. Additionally, it may be worth exploring self-supervised methods as they could potentially benefit the data hungry Transformer.

REFERENCES

[1] Isensee, F., Jaeger, P. F., Kohl, S. A., Petersen, J., and Maier-Hein, K. H., "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods* (2021).

[2] Zhou, H.-Y., Guo, J., Zhang, Y., Yu, L., Wang, L., and Yu, Y., "nnformer: Interleaved transformer for volumetric segmentation," *arXiv preprint arXiv:2109.03201* (2021).

[3] Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H. R., and Xu, D., "UNETR: Transformers for 3d medical image segmentation," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], (2022).

[4] Sobirov, I., Nazarov, O., Alasmawi, H., and Yaqub, M., "Automatic segmentation of head and neck tumor: How powerful transformers are?," in [*Medical Imaging with Deep Learning*], (2022).

[5] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S., "A convnet for the 2020s," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], (2022).

[6] Yu, W., Luo, M., Zhou, P., Si, C., Zhou, Y., Wang, X., Feng, J., and Yan, S., “Metaformer is actually what you need for vision,” in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], (2022).