Article Review: Analyzing components of a transformer under different dataset scales in 3D prostate CT segmentation

Matteo MARENGO & Manal MEFTAH - Group 14 - ENS Paris-Saclay - MVA

March 18, 2024

1 Paper Summary

1.1 Unveiling the Core Subject of the Study

The authors in [8] propose that the segmentation of tumors and organs-at-risk (OAR) in computed tomography (CT) scans could be significantly enhanced by recent deep learning innovations, particularly by the Transformer model introduced by Vaswani et al. [9]. Nonetheless, when it comes to hybrid methodologies that integrate convolutional layers with Transformer blocks, **there is a notable gap in the literature regarding the examination of variable data scales and the nuanced interplay of different architectural components.** This study offers a nuanced analysis, juxtaposing the nnFormer with variable scale datasets and experiments where blocks of the architecture are assessed.It sheds light on both the promise and the constraints of Transformer-based approaches.

1.2 Positioning with respect to the state-of-the-art

The prevailing benchmark for medical image segmentation has been the convolution-based UNet architectures [7]. However, **Could vision transformers (ViT) [2] surpass the performance of existing architectures?** While models like UNETR [3] have emerged, the authors have opted to study a hybrid model, the nnFormer [11], based on the frame of nnUnet [4], that combines shifted-window (Swin) Transformer blocks [5] with traditional convolutional frameworks. Although prior research has tried to contrast convolutional and Transformer-based methods [6], these comparisons often fall short in their assessments, not being tedious enough for a comprehensive evaluation.

1.3 Main contributions of the paper

The primary research gap identified is the exploration of the nuanced performance of hybrid Transformer networks across **varying dataset scales in medical image segmentation**, specifically prostate CT scans. While existing studies have delved into Transformers' capabilities, **there's a lack of comprehensive analysis on how different components within these networks perform under the constraints of dataset size variability**. This gap highlights the need for further investigation into the scalability of these models and their component-wise efficiency to enhance their practical applicability.

1.4 Methodology

Three experiments are made to compare different data scales and different blocks of the nnFormer **(1st: replace the Swin-Transformer block with a convolution block, 2nd: replace the self-attention mechanism with the pooling operation, 3rd: evaluate positional encoding).** To evaluate the results, metrics used are Dice [1] (similarity metric between predicted and actual mask) and the 95^{th} percentile Hausdorff Distance (HD95). To evaluate different data scales 6 different combinations were made.

1.5 Results

Experiments	**Results**
Method 1: Replacing Swin-Transformer block with convolution block	Convolution blocks surpasses Swin blocks as the size of the data increases. It is because with small dataset Conv layers do not capture semantic features. Swin blocks are not robust.
Method 2: Replacing the Self-Attention with Pooling	Max-pooling is better with small dataset scales. As self-attention mechanism overfit.
Method 3: Evaluating Positional Encoding	Conv layers have a positional inductive bias that might allow making positional encoding not effective.

Table 1: Example table with text wrapping

The authors conclude by saying that Swin-Tranformer blocks do not have advantages compared to convolutional blocks. This is mainly because Swin blocks have a low understanding of positional information.

2 Critical assessment

2.1 Strenghts & Weaknesses of the paper

Authors have lead comprehensive experiments with detailed protocols so that nnFormer can be evaluated on **different data scales and that each blocks can be fully** evaluated. Furthermore, they compare the models with same number of parameters being fair. However, they lack on the diversity of the data. Indeed, in other papers [11] more datasets are compared and not only CT scans but also MRI. Additionally, the study does not address the potential for model overfitting in scenarios of highly variable data, nor does it explore the impact of data augmentation techniques on enhancing the model's generalization capabilities.

2.2 Recommendations for improvement

While Dice and HD95 are standard metrics, incorporating **sensitivity or Jaccard index** could offer a comprehensive performance view, essential for clinical accuracy. The finding that transformers struggle with positional information suggests exploring dynamic positional encoding, **geometric positional encoding** (spatial relationships in medical images) or attention mechanisms tailored to medical imaging could improve spatial understanding in segmentation tasks.

The study indicates convolutional layers outperform transformers in larger datasets, pointing towards the necessity of **data-efficient transformer design**s. Future research could focus on developing medical-specific data augmentation techniques or hybrid models that blend convolutional layers' local feature extraction with transformers' global context awareness. Considering the data-intensive nature of transformers, **self-supervised learning methods** using the abundance of unlabeled medical images for pre-training could address data scarcity issues, enhancing model performance with limited labeled datasets. An interesting way is also multi-modal data fusion [10] to see whether accuracy increases or not when doing the three different experiments.

3 Conclusion

In essence, this study opens avenues for improving transformer applicability in medical imaging, suggesting a focus on advanced encoding techniques, hybrid architectures, and leveraging unlabeled data through self-supervised learning for future research.

References

[1] Lee R. Dice. Measures of the amount of ecologic association between species. In *Ecology. 26 (3): 297–302.*, 1945.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[3] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger Roth, and Daguang Xu. Unetr: Transformers for 3d medical image segmentation, 2021.

[4] Kohl SAA Petersen J Maier-Hein KH. Isensee F, Jaeger PF. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. In *Nat Methods*, 2021.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[6] Christos Matsoukas, Johan Fredin Haslum, Magnus Söderberg, and Kevin Smith. Is it time to replace cnns with transformers for medical images?, 2021.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[8] Yicong Tan, Prerak Mody, Viktor van der Valk, Marius Staring, and Jan van Gemert. Analyzing components of a transformer under different dataset scales in 3D prostate CT segmentation. In Olivier Colliot and Ivana Išgum, editors, *Medical Imaging 2023: Image Processing*, volume 12464, page 1246408. International Society for Optics and Photonics, SPIE, 2023.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[10] Yifei Zhang, Désiré Sidibé, Olivier Morel, and Fabrice Mériaudeau. Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, 105:104042, 2021.

[11] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation, 2022.