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

Yicong Tan [1] Prerak Mody[2] Viktor van der Valk [2] Marius Staring [2] Jan van Gemert [2]
Students: Manal Meftah [3] Matteo Marengo [3] Group 14
[1]TU Delft [2]Leiden University Medical Center [3]ENS Paris-Saclay, Master MVA

Abstract

In the pursuit of advancing medical imaging, particularly organ segmentation within prostate CT scans, this research [6] scrutinizes the efficiency and adaptability of **hybrid Transformer networks**. By meticulously altering network components and evaluating performance across different dataset scales, the study sheds light on the intricate balance between complexity and functionality. Through comprehensive experiments involving **convolutional replacements, pooling modifications, and positional encoding adjustments**, the findings offer valuable insights into the potential and limitations of hybrid Transformers in medical imaging, proposing a paradigm shift towards optimizing component selection based on dataset characteristics.

Literature Review

State-of-the-Art

The UNet [5] has been since its creation the most used network for image segmentation. The Vision Transformer [1] and more specifically the SWIN Transformer (with shifted window) [4] success leads to experimentation of applying it to the UNet to create hybrid transformer network and performs even better organ segmentation. An innovative approach has been developed known as nnUNET [3].

Research gaps

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 in medical imaging tasks.

Research Objectives

The present study investigates the following objectives:

- **Objective 1:** Evaluate the performance of hybrid Transformer networks in organ segmentation within prostate CT scans, focusing on different components' efficiency across dataset scales.
- **Objective 2:** Assess the impact of convolutional replacements, pooling strategies, and positional encoding adjustments within these networks to identify optimal configurations for medical image segmentation.

Study Methodology

The methodology implemented by the authors included a rigorous statistical evaluation of the nnFormer [8] network's performance. The network's architecture and modifications for each experiment are illustrated in Figure 1. Swin blocks have been interleaved with Embedding blocks that are convolutional.

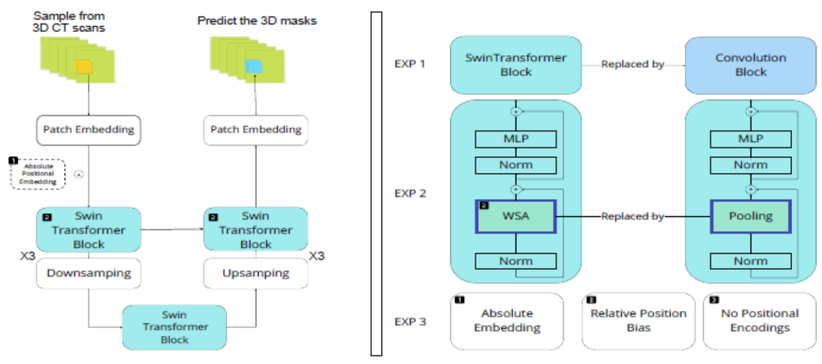

Figure 1. Network architecture of the nnFormer and the modifications for the three experiments.

The statistical approach is centered on learning curves to assess differences in segmentation accuracy across dataset scales. This method allowed for a robust analysis of the performance implications of replacing Transformer components with convolutional counterparts.

Dataset Description

The dataset comprised multi-institutional prostate CT scans used to evaluate the segmentation performance of the nnFormer model. A total of **710 CT scans** were collected from Haukeland Medical Center of Norway (HMC), Leiden University Medical Center in the Netherlands (LUMC), and Erasmus Medical Center in the Netherlands (EMC). **EMC data served as the test set, with HMC and LUMC providing training data.** Annotations for four organs—bladder, prostate, rectum, and seminal vesicles—were included.

Preprocessing involved resampling scans to median spacing, standardizing intensity ranges, and extracting $128 \times 128 \times 64$ voxel patches. The dataset variability reflected differences in clinical protocols and anatomy, with the EMC dataset presenting larger prostate and bladder volumes, posing an additional challenge in segmentation tasks.

Experiment Design and Settings

This study evaluates the nnFormer model's performance by comparing Transformer and convolutional components across varied dataset scales. The HMC and LUMC datasets, referred to as Clinic A and Clinic B respectively, form the basis of our experimental data, with A divided into subsets A1 and A2 for detailed analysis. **Six dataset combinations (A1, A2, A, A1+B, A2+B, A+B) facilitate the examination of model scalability.**

Key experimental modifications include:

- Replacing Swin-Transformer blocks with convolutions to assess local spatial feature capture.
- Substituting self-attention with pooling in smaller datasets to evaluate overfitting prevention.
- Comparing positional encoding strategies to explore their impact on model performance.

Model performance was quantified using the **Dice coefficient** and the **95th percentile Hausdorff Distance (HD95)**, with statistical significance determined via the Wilcoxon signed-rank test (p-value < 0.05). Training involved 500 epochs using a Dice and cross-entropy loss combination on an Nvidia RTX6000 GPU.

Results and Discussion

The study's outcomes highlight the nuanced performance of hybrid Transformer networks across varying dataset scales. Notably, the convolutional replacements and positional encoding adjustments were of particular interest.

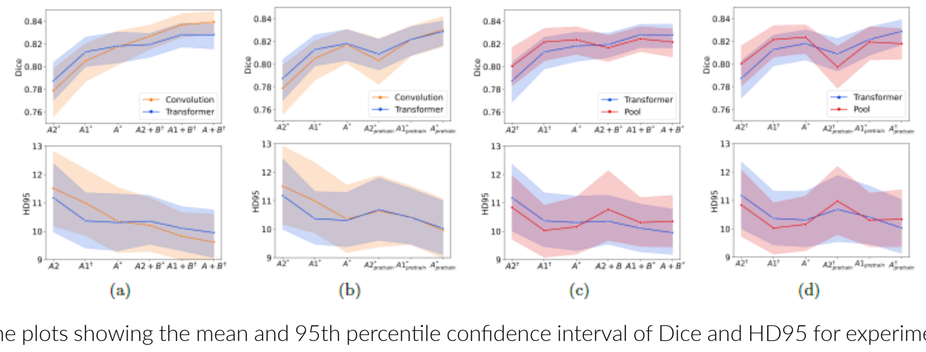

Figure 2. Line plots showing the mean and 95th percentile confidence interval of Dice and HD95 for experiments 1 and 2.

The results from Figure 2 indicate that **convolutional blocks outperform Transformer blocks as the dataset size increases**. This suggests that the local inductive biases inherent to convolutions might be more effectively capturing spatial relationships within the data at larger scales.

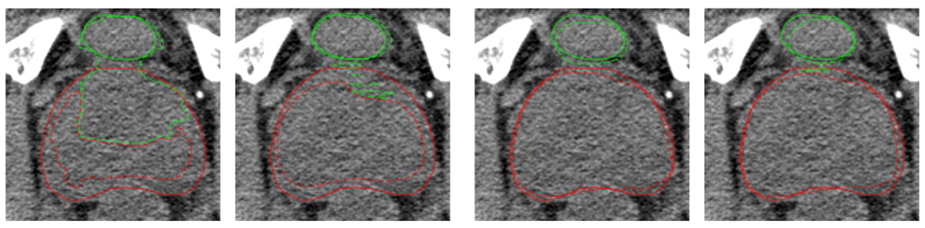

Figure 3. CT scans showing the prediction (dotted lines) and ground truth (solid line) for Self-Attention vs Pooling compared with small vs large dataset

What can be stated from Figure 2 is also that when replacing self-attention with pooling, it performs better in **larger dataset scale**. It is further confirmed in Figure 3 where the self-attention is more precise with a large dataset.

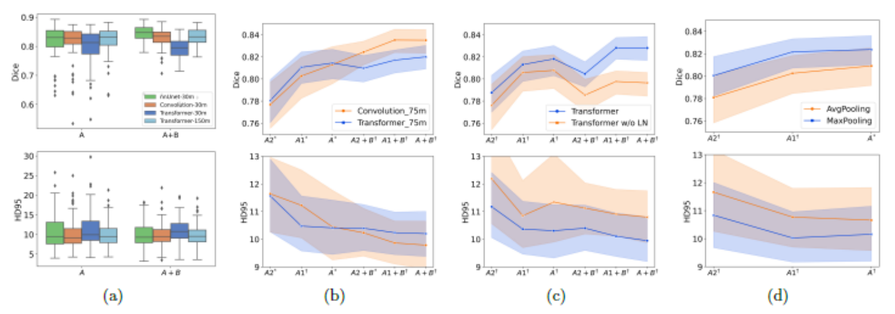

Figure 4. Ablation study outcomes showing Dice and HD95 across training scales and positional encoding methods, with statistical significance denoted.

The results in Figure 4 convey that while convolutions yield more consistent segmentation across data scales, the Transformer models can achieve comparable results without added complexity. Max-pooling has a slight advantage over avg-pooling in smaller datasets, highlighting its potential in constrained data environments.

Therefore, while convolutional blocks tend to outperform Transformer blocks in larger datasets especially, Transformer models can match the performance without complex architecture tweaks.

Critical Analysis

- One critical point is the lack of data and therefore the **generalization** issue. Indeed, the nnFormer [8] for example tested its model on MRI scans, CT scans mixing brain tumors and abdominal organs. In addition, in [6], no information are given on the patients from which the images come from.
- It can be also argued that the analysis on the positional encoding is not depth enough. The authors could have compared other encoding schemes such as **geometric positional encoding** (spatial relationships in medical images) or **dynamic positional encoding** (models could be tailored to the specific features of each image).

Conclusions

- Precise experiments to assess the network performance by removing or modifying some blocks. Different scales of dataset are evaluated.
- For the chosen dataset, Swin-Transformer block is not the best choice.
- The number of the data might be not enough for the Transformers, self-supervised learning might be a solution.
- Based on that extensive study new approaches are now being explored to study transformer blocks compared to convolutional ones [2].

References

[1] 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.

[2] Syed Nouman Hasany, Caroline Petitjean, and Fabrice Meriaudeau.
A study of attention information from transformer layers in hybrid medical image segmentation networks.
In Olivier Colliot and Ivana Išgum, editors, *Medical Imaging 2023: Image Processing*, volume 12464, page 124641O. International Society for Optics and Photonics, SPIE, 2023.

[3] 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.

[4] 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.

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

[6] 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.

[7] 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.