Comparing modern segmentation architectures under low data regime for PET-CT tumor segmentation

Gustavo Andrade-Miranda, Vincent Jaouen, Dimitris Visvikis, Pierre-Henri Conze

Comparing modern segmentation architectures under low data regime for PET-CT tumor segmentation

Gustavo Andrade-Miranda, Vincent Jaouen, Dimitris Visvikis and Pierre-Henri Conze

***Abstract*—In medical imaging, low data regimes arise from practical situations where not only data labeling but also data collection itself is expensive. Most modern segmentation architectures compensate data scarcity by the intensive use of data augmentation. However, another important lever for improved generalization power is the neural architecture itself. In this context, the goal of this paper is to provide insights into the performances of various modern hybrid and Transformers-based segmentation networks under various data regimes. We believe this comparative analysis is crucial to better understand and guide future research on segmentation scenarios suffering from data scarcity. We conducted a series of controlled studies to analyze PET-CT head and neck tumor segmentation performance of hybrid models versus pure Transformer-based models with a variable number of training subjects. One path versus multi-path encoders were also investigated to study the management of multi-modality. Results on the publicly-available HECKTOR 2021 dataset show that hybrid pipelines generally tend to outperform simple Transformers-based models and perform similarly with respect to hierarchical Transformers. In these experiments, multi-path encoders surpass the one path strategy, and no particular improvement using multi-modal Transformers was observed.**

I. Introduction

Deep learning (DL) models better generalize with more data. As a result, DL models often rely on large amounts of annotated samples to achieve good performance, especially in computer vision or natural language processing. However, there are many domains where the amount of data is limited. Particularly, medical data and annotations are usually scarce and difficult to obtain for a variety of reasons including legal regulations, rare diseases or when expert-level annotations are prohibitively expensive or hard to get.

A common strategy in scarcer data regime is to resort to *data augmentation* (DA) techniques or to limit the representational capacity of the model. In medical imaging, DA is commonly conducted by performing various small perturbations to the real data on the fly during training, such as rotations, scaling, Gaussian noise, blurring, brightness, resampling, gamma correction or mirroring [1]. Another strategy is to generate synthetic data whose distribution has to be as close as possible to the real distribution using deep adversarial networks [2].

On the other hand, the representational ability of a DL model determines its flexibility and its capacity to learn which family of functions the learning algorithm can choose to reduce a training objective. Under a low data regime, models with insufficient capacity are unable to solve complex tasks, and models with higher capacity may tend to overfit the training set. This problem is even more noteworthy in successful modern architectures such as Visual Transformers (ViT) [3] that usually rely on even larger amount of data.

Emerging transformers models can either be hybrid when used in conjunction with convolutional layers (CNN-ViT) or purely Transformers-based. In this work, we conduct a comparative quantitative analysis to study the representational capacity between a range of Transformers-based models under a low data regime for PET-CT tumor segmentation. We also investigate to what extent such models could benefit from the use of multi-modal data to exploit the complementary and redundancy of information across modalities. In this direction, the performance reached by various modern hybrid and Transformers-based architectures is compared for PET-CT head and neck tumor segmentation with increasing number of training samples. The rest of the paper is organized as follows. Sect.II introduces both ViT and CNN-ViT architectures as well as the model implementation details. Sect.III explains the evaluation strategy and presents the results we obtained. Conclusions and perspectives are given in Sect.IV.

The authors are with LaTIM UMR 1101, Inserm, Brest, France. V. Jaouen and P.-H. Conze are also with IMT Atlantique, Brest, France.

Corresponding address: gustavo-xavier.andrade-miranda@inserm.fr

II. Methods

Hybrid and Transformer-based architectures follow a U-shaped design [4] in which the extracted feature representations from encoder layers are fused with their decoder counterparts by concatenation through skip-connections. The hybrid models are made of ResNet encoding blocks at the shallower levels, while deeper levels are encoded with Transformer blocks. Given an input $X \in \mathbb{R}^{C \times H \times W \times D}$ with spatial resolution $H \times W$, D as depth dimension (# of slices) and C channels (# of modalities), downsampling blocks gradually encode input images into a low-resolution/high-level feature representation $\mathcal{F} \in \mathbb{R}^{F \times \frac{H}{8} \times \frac{W}{8} \times \frac{D}{8}}$ where F represents the number of encoded feature maps. Such backbone can be extended to a multi-encoder-based framework, where independent encoders learn modality-specific feature representations. Then, we distinguish two types of Transformer blocks: vanilla ViT and single-stream ViT. The vanilla ViT block (ViT_v) follows the design proposed in the original ViT paper [3]. Meanwhile, single-stream ViT (ViT_s) layers collectively operate both images modalities to enable early and unconstrained fusion of cross-modal information. No extra modal-type embedding or learnable token-position are considered. On the other hand, we consider two Transformers-based models, UNETR [5] and Swin UNETR [6], which both work with 1D sequences. However, the way features are extracted differs. UNETR is solely built on a plain backbone (no hierarchical ViT) whereas Swin UNETR extracts features at several resolutions (hierarchical ViT), making it more suitable for dense prediction tasks as image segmentation.

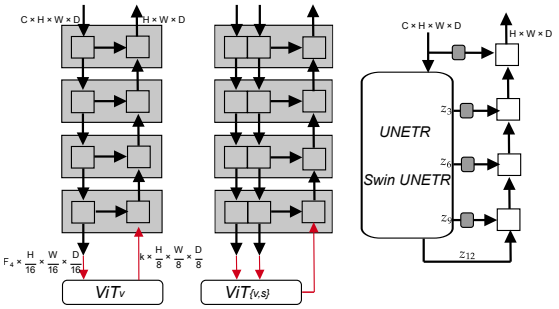


Fig. 1: Left: CNN-ViT; Middle: Multi-path CNN-ViT; Right: Transformers-based models.

The proposed hybrid-based models are denoted as UNETR, Swin UNETR, CNN+ViT$_v$-B/P, MCNN+ViT$_v$-B/P, and MCNN+ViT$_s$-B/P where the subscripts v and s represent the type of ViT block, B indicates the base ViT model configuration, P the patch size and M stands for multi-encoder based CNN. The different models are illustrated in Fig.1.

III. Experiments and Results

We performed our experiments on the HECKTOR 2021 dataset for PET-CT head and neck tumor segmentation [7]. The training set comprises 224 cases from 5 clinical centers. All models were trained using Nvidia A6000 and TitanRTX GPUs. All models were trained for a total of $10,000$ iterations, with a batch size of 2 and NovoGrad as optimizer (learning rate $=$ 0.001). The training objective was the sum of Dice and cross-entropy losses. The CNN feature maps F were set to 16, 32, 64 and 128 for all experiments. We used different patch resolutions as inputs to the Transformer blocks depending on the model implemented. For hybrid models, we used a patch size P of $1 \times 1 \times 1$. Meanwhile, the patch size was set to $16 \times 16 \times 16$ for UNETR and $4 \times 4 \times 4$ for Swin UNETR. The employed Transformer blocks followed the ViT base configuration [3]. We did not use any pre-trained weights, neither for the CNN nor the ViT blocks. For a fair comparison, we followed both pre-processing and data augmentation strategies used in nnU-Net [1] for all implemented models. To evaluate how modern segmentation architectures react under a low data regime, we progressively scaled up the size of our training dataset while keeping the test set constant (80 samples). We construct a total of ten subsets starting from 10 samples until we reach 100 samples (step = 10). Over each new training set, we evaluated the performance of our models using averaged Dice (DSC) scores. Results obtained for each subset are depicted in Fig.2 meanwhile Fig.3 shows qualitative segmentation results for the best hybrid and Transformer-based models over the test set.

The present results suggest that hybrid methods perform better than pure hierarchical Transformers-based architectures (UNETR). In addition, we observed that multi-path CNN-ViT outperformed the one path CNN-ViT. It could be due to the fact that these architectures facilitate the learning of inter-and intra-modal interactions (inter- at single-stream level and intra- at Transformer levels). We also found that no significant difference could be established between ViT$_v$ and ViT$_s$ in hybrid models. In addition, we observed that excluding

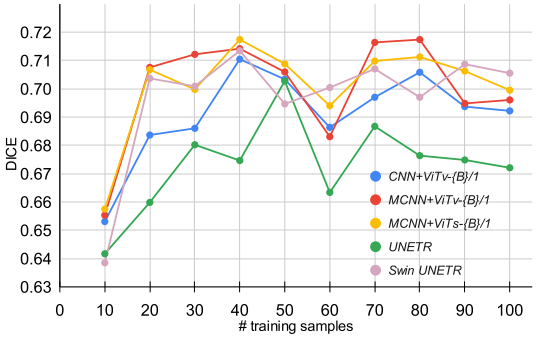
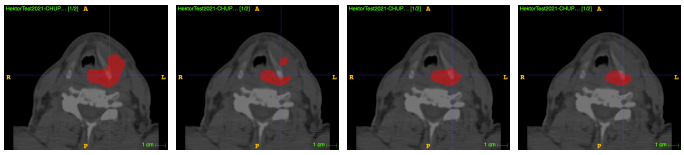


Fig. 2: Dice scores reached by hybrid CNN-ViT and ViT segmentation models for a variable number of training subjects.

positional embedding and modal embedding did not affect the performance of MCNN+ViT$_s$-B/P. Multi-path outperforms all models in very low data regime (less than 30 samples). This could be linked to the fact that convolution introduces appropriate inductive bias that helps to learn strong representations under small data regimes. This finding agrees with the current state of the art which claims that no hirarchical Transformers (UNETR) has poor performance in small data regimes. Lastly, Swin transformer improves UNETR thanks to its hierarchical architecture which is computed using shifted windows.



(a) (b) (c) (d)

Fig. 3: Visual results for 10 training samples only: (*a*) ground truth, (*b*) CNN+ViT$_v$, (*c*) MCNN+ViT$_v$ and (*d*) UNETR.

IV. Conclusion

In this work, we compared a variety of ViT and hybrid CNN-ViT architectures in a context of low data regime for PET-CT tumor segmentation. This analysis will be extended to additional multi-modal datasets to provide a deeper understanding of how such deep models behave when facing data scarcity.

References

[1] Isensee *et al.*, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, 2020.

[2] Q. Li *et al.*, "TumorGAN: A multi-modal data augmentation framework for brain tumor segmentation," *Sensors*, 2020.

[3] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015.

[5] A. Hatamizadeh *et al.*, "UNETR: Transformers for 3D medical image segmentation," in *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 574–584.

[6] A. Hatamizadeh *et al.*, "Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images," *arXiv preprint arXiv:2201.01266*, 2022.

[7] V. Andrearczyk *et al.*, "Overview of the HECKTOR challenge at MICCAI 2021: Automatic head and neck tumor segmentation and outcome prediction in PET/CT images," in *Head and Neck Tumor Segmentation and Outcome Prediction*, 2022.