DDViT: Double-Level Fusion Domain Adapter Vision Transformer (Student Abstract)

Linpeng Sun, Victor S. Sheng

Department of Computer Science, Texas Tech University
linsun@ttu.edu, victor.sheng@ttu.edu

Abstract

With the help of Vision transformers (ViTs), medical image segmentation was able to achieve outstanding performance. In particular, they overcome the limitation of convolutional neural networks (CNNs) which rely on local receptive fields. ViTs use self-attention mechanisms to consider relationships between all image pixels or patches simultaneously. However, they require large datasets for training and did not perform well on capturing low-level features. To that end, we propose DDViT, a novel ViT model that unites a CNN to alleviate data-hunger for medical image segmentation with two multi-scale feature representations. Significantly, our approach incorporates a ViT with a plug-in domain adapter (DA) with Double-Level Fusion (DLF) technique, complemented by a mutual knowledge distillation paradigm, facilitating the seamless exchange of knowledge between a universal network and specialized domain-specific network branches. The DLF framework plays a pivotal role in our encoder-decoder architecture, combining the innovation of the TransFuse module with a robust CNN-based encoder. Extensive experimentation across diverse medical image segmentation datasets underscores the remarkable efficacy of DDViT when compared to alternative approaches based on CNNs and Transformer-based models.

Introduction

Medical imaging segmentation (MIS) plays a pivotal role in diagnosing and treating numerous health conditions, ensuring precise and targeted interventions. However, the ideal segmentation tool must overcome the challenge of data-hunger to be applicable across a wider range of scenarios. Traditional CNNs, despite their remarkable efficacy, face inherent limitations. Their constrained receptive field and intrinsic inductive biases often hinder them from grasping global contexts and long-range dependencies, thus potentially impacting their overall performance.

ViTs' innovative neural architectures harness the power of multi-head self-attention mechanisms, enabling them to adeptly capture long-range relationships and global contexts (Zhang, Liu, and Hu 2021). However, the transition to ViTs isn't without its set of challenges. Although they rival CNNs in performance, ViTs demand vast amounts of data and are computationally intensive. Recognizing these challenges, the research community introduced solutions such as the DeiT, Swin Transformer, and pyramid vision transformer to mitigate these concerns.

Furthermore, recent advancements in multi-scale feature representations, as seen in models like CrossViT (Chen, Fan, and Panda 2021) and DS-TransUNet (Lin et al. 2021), have demonstrated impressive results, particularly in the realm of medical image segmentation. Yet, a persistent challenge remains: Vision transformers, while powerful, often overlook some critical low-level features. This observation has spurred the development of hybrid models like HiFormer (Heidari et al. 2023) and TransUnet, which aim to merge the strengths of both CNNs and transformers. However, seamlessly blending features and leveraging multi-scale information remains a complex task for these hybrid models.

In this evolving landscape, the introduction of the domain adapter by MDViT (Du et al. 2023) offers a promising direction. By adaptively harnessing knowledge across multiple minor data sources or domains, the domain adapter addresses the data-hunger challenge and counteracts negative knowledge transfer. With such innovations, the future of medical imaging segmentation is poised for more accurate, efficient, and wide-ranging applications.

Approach

Encoder

Our proposed encoder is composed of two hierarchical models, CNN and TransFuse+DA, with the DLF module that enriches the retrieved features and prepares them to be fed into the decoder. Since using CNNs or transformers separately causes either local or global features to be neglected, which affects the model's performance, we first utilize the CNN locality trait to obtain local features. Here, the CNN and TransFuse each include three distinct levels. We transfer local features of each level to the corresponding TransFuse's level via a skip connection to attain universal representations. Then each transferred CNN level is added with its parallel transformer level and passes through a Patch Merging module to produce a hierarchical representation. We exploit the hierarchical design to take advantage of multi-scale representations. The largest and smallest levels go into the DLF module to exchange information from different scales and

generate more powerful features.

The feature extractor if the CNN module which takes an input image $X \in R^{H \times W \times C}$ with spatial dimensions H and W, and C channels, it is first fed into the CNN module. CNN module consists of three levels, from which a skip connection is connected to the associated transformer's level using a Conv 1×1 to compensate for low-level missing information of transformers and recover localized spatial information.

TransFuse is designed with two parallel branches that process information in distinctive ways. The first branch is a CNN branch that starts by focusing on local features and gradually expands its receptive field, thereby encoding information from local to global. In contrast, the second branch, a Transformer branch, begins its operations with a global self-attention mechanism, ultimately narrowing down to recover fine local details by the end of its processing.

Given the parallel structure of TransFuse, especially within the Transformer branch, the introduction of the "domain adapter" as a plug-in is particularly compelling. Instead of introducing additional domain-specific layers, the domain adapter leverages the inherent multi-head self-attention structure in the Transformer branch for domain adaptation. To integrate domain-specific insights, the domain adapter follows two key processes:

Attention Generation: A domain label vector is transformed into a domain-aware vector, which in turn helps produce attention for each head within the Transformer branch.

Information Selection: After obtaining the features from each head, the generated attention values are used to fine-tune or modulate the information based on its relevance to the domain.

The fusion of information across domains is then seamlessly handled by the BiFusion Module. This innovative module, combined with domain adaptation capabilities, results in a powerful mechanism that selectively fuses information drawn from multi-level feature maps. This fused data is then harnessed to produce segmentation results through a series of gated skip-connections.

By integrating the domain adapter, TransFuse not only remains sensitive to low-level contexts but also efficiently captures domain-specific nuances, making its resultant representation both robust and compact across multiple domains.

Decoder

Driven by the principles of Semantic Feature Pyramid Networks, we craft a decoder that blends features from which takes the smallest (P^s) and largest (P^l) CNN levels and employs a cross-attention mechanism to fuse information across scales to form a cohesive mask feature. Initially, the DLF module supplies the low and high-resolution feature maps, P^s and P^l. P^s, with dimensions (H/16, W/16), goes through a ConvUp block. This block executes two phases of 3×3 Conv, $2\times$ bilinear upsampling, Group Norm, and ReLU, adjusting its resolution to (H/4, W/4). On the other hand, P^l, already at a resolution of (H/4, W/4), undergoes a Conv Block. This block uses a 3×3 Conv, Group Norm, and ReLU, maintaining its resolution. The combined outputs of the processed P^s and P^l then route through another ConvUp block to produce a comprehensive $H \times W$ feature map. After channeling this feature map through a 3×3 Conv in the segmentation head, the conclusive segmentation map forms.

	Seg-Net	U-Net	TransFuse	DDViT
Normal	**0.777**	0.411	0.554	0.725
Polyp	0.937	**0.965**	0.871	0.904
High-grade IN	0.894	0.895	0.812	**0.902**
Low-grade IN	0.924	0.911	0.864	**0.937**
Adenocarcinoma	0.865	0.887	0.834	**0.901**
Serrated adenoma	0.907	0.938	0.855	**0.951**

Table 1: Comparison DiceRatio results of the proposed method on the EBHI-Seg dataset.

Experiment and Results

We evaluate our DDViT model on Enteroscope Biopsy Histopathological Hematoxylin and Eosin Image Dataset for Image Segmentation Tasks (EBHI-Seg) (Shi et al. 2022), which is also known as Colorectal cancer. It is a prevalent and potentially life-threatening disease, remains a significant public health concern worldwide. The successful management and treatment of colorectal cancer are profoundly linked to the timing of its diagnosis. Early detection can significantly improving the chances of a positive outcome. This dataset contained 4,452 images of six types of tumor differentiation stages and the corresponding ground truth images. We use 890 dialogues for training, 890 for validation, and the remaining for testing. Our experimental comparison mainly includes several CNN-based MIS models and ViT models, Seg-Net, U-Net and TransFuse. We use the Dice ratio metric, a standard metric used in medical images that is often utilized to evaluate the performance of image segmentation algorithms, as evaluation indicator to evaluate the performance of DDViT. The experimental results are shown in Table 1. Overall, DDViT outperforms conventional deep learning methods such as Seg-Net and U-net to a modest extent. Furthermore, its efficacy surpasses that of TransFuse, particularly notable when dealing with datasets of limited scale.

Conclusion

In this paper, we introduce DDViT, a model designed to counteract the challenges associated with ViTs, namely their insatiable data needs and their subpar performance on small datasets, especially concerning low-level feature capture. We meld the local representations sourced from a CNN-based encoder with the global insights drawn from a ViT module enhanced with domain adapters, aiming for bridges the gap between local and global feature representations. When tested on the EBHI dataset, not only does DDViT surpass traditional CNN-based MIS models, but it also excels in preserving intricate low-level features and effectively modeling extended-range interactions. The model's ability to synergize the strengths of CNNs and ViTs, coupled with the unique features of the DLF framework, positions DDViT as a promising avenue for advancing computer-aided medical diagnosis and imaging applications.

References

Chen, C.-F.; Fan, Q.; and Panda, R. 2021. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification. arXiv:2103.14899.

Du, S.; Bayasi, N.; Harmarneh, G.; and Garbi, R. 2023. MD-ViT: Multi-domain Vision Transformer for Small Medical Image Segmentation Datasets. arXiv:2307.02100.

Heidari, M.; Kazerouni, A.; Soltany, M.; Azad, R.; Aghdam, E. K.; Cohen-Adad, J.; and Merhof, D. 2023. HiFormer: Hierarchical Multi-scale Representations Using Transformers for Medical Image Segmentation. arXiv:2207.08518.

Lin, A.; Chen, B.; Xu, J.; Zhang, Z.; and Lu, G. 2021. DS-TransUNet:Dual Swin Transformer U-Net for Medical Image Segmentation. arXiv:2106.06716.

Shi, L.; Li, X.; Hu, W.; Chen, H.; Chen, J.; Fan, Z.; Gao, M.; Jing, Y.; Lu, G.; Ma, D.; Ma, Z.; Meng, Q.; Tang, D.; Sun, H.; Grzegorzek, M.; Qi, S.; Teng, Y.; and Li, C. 2022. EBHI-Seg: A Novel Enteroscope Biopsy Histopathological Haematoxylin and Eosin Image Dataset for Image Segmentation Tasks. arXiv:2212.00532.

Zhang, Y.; Liu, H.; and Hu, Q. 2021. TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation. arXiv:2102.08005.