Can Frozen Transformers in Large Language Models Help with Medical Image Segmentation?

Juntao Jiang[1] JUNTAOJIANG@ZJU.EDU.CN
[1] *Zhejiang University, 38 Zheda Road, Hangzhou, China*
Yong Liu*[1] YONGLIU@IIPC.ZJU.EDU.CN

Editors: Accepted for publication at MIDL 2024

Abstract

Transformer models shine in medical image segmentation by harnessing their self-attention mechanism to capture global information, thus boosting segmentation accuracy. Recent research has unveiled that large language models (LLMs), trained solely on text, surprisingly excel at visual tasks even without language, through a simple strategy: integrating a frozen transformer block from pre-trained LLMs as a direct visual token processor. This paper applies this approach to medical image segmentation by combining frozen transformer blocks with TransUNet. Experiments are conducted on BTCV, ACDC, ISIC 2017, CVC-ClinicDB, CVC-ColonDB and BUSI datasets, demonstrating some improvements compared to the baseline. The code will be released at: https://github.com/juntaoJianggavin/LLM4MedSeg.

Keywords: Medical image segmentation, Large Language Models, Frozen transformers, TransUNet

1. Introduction

Medical image segmentation refers to the process of partitioning medical images into meaningful regions or structures. By accurately delineating organs, tissues, tumors, and other anatomical structures with designed algorithms automatically, computer-assisted segmentation facilitates quantitative analysis, patient-specific modeling, and personalized healthcare delivery. UNet(Ronneberger et al., 2015) and its variants hold a prominent position in deep learning-based medical image segmentation due to its combination of encoder-decoder structure and skip connections, effectively capturing multiscale features. Transformer-based UNet variants such as TransUNet (Chen et al., 2021), Swin-UNet (Cao et al., 2022) and Ds-transunet (Lin et al., 2022) exhibit remarkable capabilities in medical image segmentation owing to their overall architectures with self-attention mechanism, which adeptly captures global information and have stronger visual encoding capabilities.

Recent research (Pang et al., 2023) unveils the surprising effectiveness of large language models (LLMs), trained solely on text, as potent encoders for pure visual tasks by leveraging a frozen transformer block from pre-trained LLMs and another research (Lai et al., 2024) show that the LLM transformer block with global residual connections can boost medical image classification. This paper applies the method in (Pang et al., 2023) for medical image segmentation tasks, adding a frozen transformer block from LLaMA-7B to the encoder of TransUNet. Experiments on BTCV (Landman et al., 2015), ACDC (Bernard et al., 2018), ISIC 2017 (Codella et al., 2018), CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Bernal et al., 2012) and BUSI (Al-Dhabyani et al., 2020) datasets show that this method can lead to some improvements.

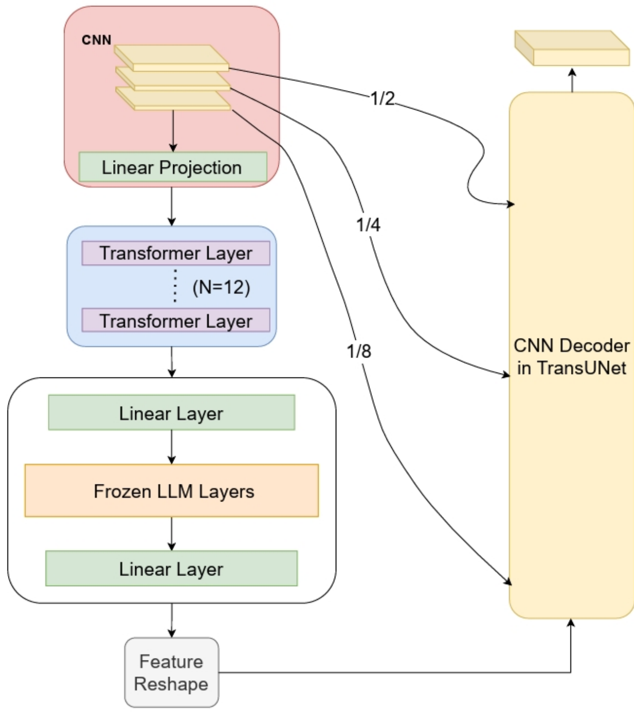

Figure 1: The architecture of TransUNet with frozen Transformer blocks in LLMs (LLM-TransUNet)

2. Method

The architecture of TransUNet with frozen Transformer blocks in LLMs (LLM-TransUNet) can be seen in Figure 1. A frozen transformer block from pre-trained LLaMA-7B (Touvron et al., 2023) is inserted after Transformer layers in TransUNet. As the feature dimensions are different between the transformers in TransUNet and the language model, two linear layers are used to align the dimensionality like the design in (Pang et al., 2023).

3. Experiments

3.1. Implementation Details

All experiments are conducted on Tesla PG500-216 GPU platform with 32G RAM. The resolution of input images is 256×256. For ISIC 2017 (Codella et al., 2018), CVC-ClinicDB (Bernal et al., 2015), CVC-ColonDB (Bernal et al., 2012) and BUSI (Al-Dhabyani et al., 2020), the total training epochs are 200, and the batch size in training, validation and testing is 8. The optimizer used is ADAM (Kingma and Ba, 2017). The initial learning rate is 0.0001 while the minimum learning rate is 0.00001. A CosineAnnealingLR (Loshchilov and Hutter, 2016) scheduler is utilized. Rotation by 90 degrees clockwise for random times, random flipping and normalization methods are applied for data processing and augmentation. The evaluation metrics in validation are *IOU* of the lesions and in testing are *IOU* and *Dice* of the lesions. The loss function used is a mixed loss combining binary cross entropy loss and dice loss (Milletari et al., 2016):$\mathcal{L} = 0.5BCE(\hat{y}, y) + Dice(\hat{y}, y)$. For CVC-ClinicDB, CVC-ColonDB and BUSI, the testing sets are split from datasets with a ratio of 0.2. Then the validation sets are split from the training set with a ratio of 0.2. The random states for splitting are all 41. ISIC 2017 has its own validation and testing set.

For BTCV (Landman et al., 2015) and ACDC (Bernard et al., 2018), all settings except for the device specification and the input size of images, including the dataset preprocessing methods and training details, are all aligned with the configuration detailed in (Chen et al., 2021).

3.2. Results

The experimental results are shown in Table 1 and Table 2, which show that a frozen transformer layer in LLM can lead to improvement in medical image segmentation tasks. The mDice refers to mean Dice and the mHD95 refers to mean Hausdorff Distance 95. The explanation may be as the hypothesis in (Pang et al., 2023) mentioned – the incorporated LLM blocks may distinguish the informative tokens and amplify their effect.

Table 1: Comparison Experimental Results on ISIC 2017, BUSI, CVC-ClinicDB and CVC-ColonDB datasets

Methods	ISIC 2017		BUSI		CVC-ClinicDB		CVC-ColonDB	
	IoU	Dice	IoU	Dice	IoU	Dice	IoU	Dice
TransUNet	0.7355	0.8442	0.6323	0.7634	0.8500	0.9183	0.8526	0.9182
LLM-TransUNet	**0.7532**	**0.8560**	**0.6436**	**0.7742**	**0.8525**	**0.9184**	**0.8728**	**0.9311**

Table 2: Comparison Experimental Results on BTCV and ACDC datasets

Methods	BTCV		ACDC	
	mDice	mHD95	mDice	mHD95
TransUNet	0.7845	30.4188	0.8988	2.6160
LLM-TransUNet	**0.7869**	**25.7007**	**0.9029**	**1.2210**

4. Conclusion

This paper explores integrating frozen transformer blocks from large language models (LLMs) into the Trans-UNet architecture for medical image segmentation. Experimental results demonstrate some improvements compared to the baseline, further proving that the LLM layers can be useful in visual recognition tasks. The designed module has the potential to become a plug-and-play component in medical image segmentation tasks. However, some datasets are quite small, thus the way of splitting datasets has a significant impact on final results. More experiments are needed to demonstrate the blocks can truly lead to improvements. Also, more baselines like Swin-UNet and Ds-transunet can be used for experiments to further explore this topic.

References

Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. *Data in brief*, 28:104863, 2020.

Jorge Bernal, Javier Sánchez, and Fernando Vilarino. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition*, 45(9):3166–3182, 2012.

Jorge Bernal, F Javier Sánchez, Gloria Fernández-Esparrach, Debora Gil, Cristina Rodríguez, and Fernando Vilariño. Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. *Computerized medical imaging and graphics*, 43:99–111, 2015.

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11): 2514–2525, 2018.

Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Zhixin Lai, Jing Wu, Suiyao Chen, Yucheng Zhou, and Naira Hovakimyan. Residual-based language models are free boosters for biomedical imaging, 2024.

Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, volume 5, page 12, 2015.

Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. Ds-transunet: Dual swin transformer u-net for medical image segmentation. *IEEE Transactions on Instrumentation and Measurement*, 71:1–15, 2022.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016.

Ziqi Pang, Ziyang Xie, Yunze Man, and Yu-Xiong Wang. Frozen transformers in language models are effective visual encoder layers, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.