Exploring Language Grounding for Few-Shot Learning in 3D Medical Segmentation

Mohammadreza Dorkhah
Simon Fraser University
mda84@sfu.ca

Yide Ma
Simon Fraser University
yma73@sfu.ca

April 8, 2023

Abstract

In this paper, we propose a novel approach for medical image segmentation by introducing top-down language grounding to the current state-of-the-art model on the Medical Segmentation Decathlon dataset[1]. Our method leverages language input to guide the segmentation process, enabling the model to focus on specific objects or regions of interest. We evaluate our method on the MSD dataset [1] for multi-organ segmentation and demonstrate improved performance compared to traditional bottom-up segmentation methods. Our results show that language grounding can effectively facilitate the medical image segmentation task, allowing for more precise and efficient segmentation results.

Keywords

Multi-modal Machine Learning, 3D Medical Image Segmentation, Brain Tumor Segmentation, Swin Transformer

1 Introduction

Medical images are critical for medical diagnosis and treatment, but processing and collecting these images can be time-consuming and challenging. To address this issue, we propose to integrate language grounding into the medical image segmentation process to facilitate and improve the performance of the model. Our objective is to investigate the impact of language grounding on the top-down medical image segmentation task, and to evaluate the zero-shot capacity of our proposed method. We aim to demonstrate that by incorporating language grounding, we can achieve improved segmentation performance, providing a more efficient and effective approach for medical image processing. In our proposed model, the input is the language description you are aiming to segment and the body part 3D scans from CT or MRI medical images, the output would be segmented 3D scans based on your description.

1.1 Motivation

The motivation behind our paper stems from the observation that there are significant similarities between different organs in the human body, particularly between the liver and liver tumors, and between the left and right kidneys. One possible explanation for these similarities could be the presence of common words or features across these organs. However, there are also differences in the relationships between certain organs, such as the proximity of the hepatic vessel to the liver versus the kidneys, or the location of the adrenal glands closer to the kidneys. By using clip[4]embedding, we were able to identify semantic relationships between these organs, which could be leveraged to enhance segmentation and enable the model to segment new classes of organs. Our motivation for this work is to develop more accurate and robust segmentation models that can better differentiate between different organs and improve the overall quality of medical imaging analysis.

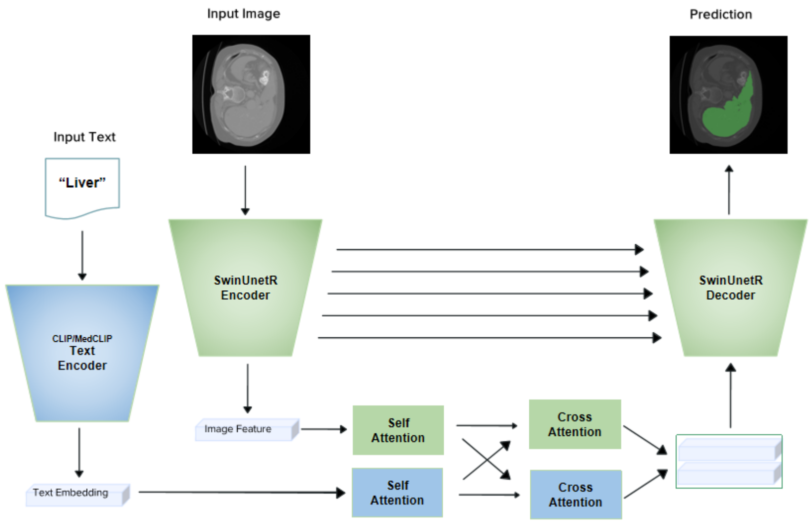

Figure 1: We have developed a new model by incorporating a fused module with both self-attention and cross-attention mechanisms into the state-of-the-art 3D medical segmentation model Swin-UNETR[2]. Our goal in doing so is to enhance the model's robustness by introducing language features into the latent space. This, we believe, will enable the model to perform zero-shot and few-shot tasks with greater accuracy and efficiency.

2 Related Work

2.1 Swin-UNETR

Since the popularity of transformers-based deep learning models, more and more 3D medical image segmentation models start to use transformers. For the MSD dataset [1] , currently the state-of-the-art model in the benchmark is the Swin-UNETR [2]. It uses Swin transformers [3] blocks and use it in a U-net structure.[7] We will refine the Swin-UNETR [2] model structure by introduce language grounding to the bottleneck feature. For modality fusion between language and 3D scans, language grounded 3D indoor segmentation [6] proposed an idea of projecting different modality to the same latent vector space.

2.2 MedCLIP

MedCLIP [8] is a state-of-the-art language model specifically trained on clinical text data, including electronic health records, radiology reports, and pathology reports. It is based on the popular transformer architecture and is pre-trained on a large corpus of clinical text data, followed by fine-tuning on downstream tasks such as named entity recognition, relation extraction, and prediction of clinical outcomes. MedCLIP has demonstrated high accuracy and outperformed other language models on several clinical NLP tasks, making it a valuable tool for healthcare research and clinical practice.

3 Approach

The model we propose builds upon the state-of-the-art Swin-UNETR architecture[2], which combines Swin transformers[3] as blocks and a Unet-based [5] structure. We believe that the bottleneck feature located at the bottom of the Swin-UNETR architecture [2] represents the latent space. This architecture can be regarded as having an encoder on the left side and a decoder on the right side. Our proposed model involves adding a language grounding module to the bottleneck feature and fusing it

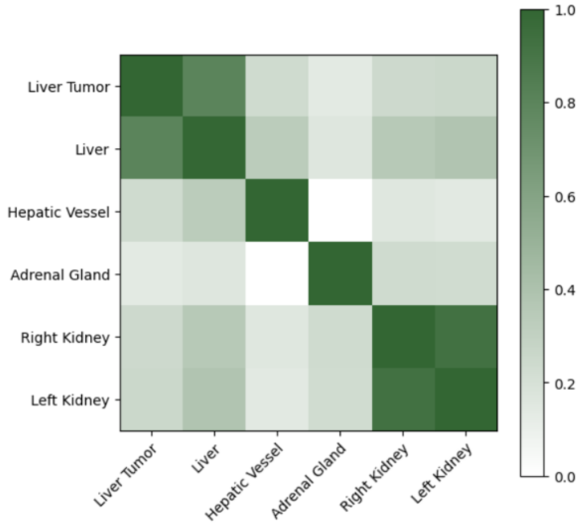

Figure 2: This similarity matrix in calculated by the inner product of different embedding. We can observe the score represent relationships between same organs and different organs. It also shows relationship between a organ and the tumour of that organ.

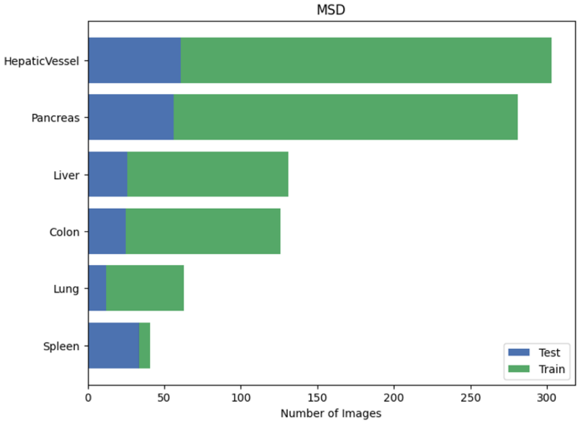

Figure 3: Organs Distribution

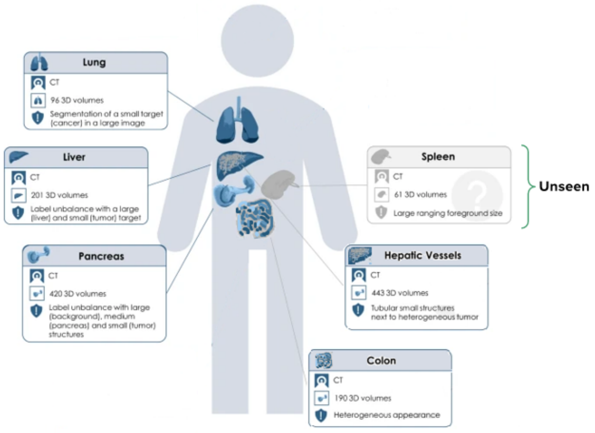
Lung
CT
96 3D volumes
Segmentation of a small target (cancer) in a large image
Liver
CT
201 3D volumes
Label unbalance with a large (liver) and small (tumor) target
Pancreas
CT
420 3D volumes
Label unbalance with large (background), medium (pancreas) and small (tumor) structures
Spleen
CT
61 3D volumes
Large ranging foreground size
Unseen
Hepatic Vessels
CT
443 3D volumes
Tubular small structures next to heterogeneous tumor
Colon
CT
190 3D volumes
Heterogeneous appearance

Figure 4: Organs Distribution

into the decoder. The fusion module includes two layers, a self-attention and a cross-attention layer. The reason for only using two layers is because our dataset did not increase. We found adding more layers could make the model overfitting. In addition, since we have both self-attention and cross-attention modules, we think both of them help the fusion part. In the training phases, our med-clip language stream encoder is locked, and we fine-tune the Swin-UNETR[2] [2] and train the fusion module. We have conducted experiments with different types of fusion modules and found that this single structure outperforms others in zero-shot and few-shot scenarios. Our model represents an innovative approach to combining language and image processing, and we believe it has significant potential for a range of applications. Our proposed model is shown as Figure 1.

4 Experiments

4.1 Dataset

The datasets we used is MSD [1]. It is a benchmark dataset for evaluating and comparing the performance

Table 1: Experimental results

Different Embedding	Seen Average	Unseen Spleen
Random	0.4046	0.0000
One-hot	0.3826	0.0000
CLIP[4]	0.3845	0.0001
MedCLIP[8]	0.2185	0.0369
CLIP[4] Prompt	0.3561	0.0022
CLIP[4] Attention	0.4166	**0.4895**

of medical image segmentation algorithms. It consists of 10 different medical image segmentation tasks covering various anatomical structures, imaging modalities, and pathologies. The dataset contains 3D and 4D CT scans and MRIs with corresponding manual segmentation masks, split into training, validation, and test sets. Since we are mainly doing experiments on zero-shot and few-shot. We used only CT (Computed Tomography) images (947 images) belonging to the Liver, Lung, Pancreas, Colon, Hepatic Vessel, and Spleen for training and testing the model. The Spleen was excluded from the training set and reserved for zero-shot and few-shot experiments.

4.2 Metric

The metric for evaluation purpose is the dice score. The dice score is the measurement of overlap between two sets of data and ranges from 0 to 1.

$$Dice(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}$$

4.3 Analysis

Based on the data presented in the table, it is clear that all types of embeddings have lower averages compared to a single Swin-UNETR[2] model. This could be due to the fact that adding language grounding may not necessarily improve the segmentation task itself. Additionally, the fusion module convention between the two feature represent two streams latent space may actually be detrimental to the segmentation tasks when language grounding is added. However, when it comes to zero-shot/few-shot tasks, the addition of language grounding outperforms all types of embeddings. This is a significant improvement as it allows the model to have zero-shot/few-shot capacity for the first time.

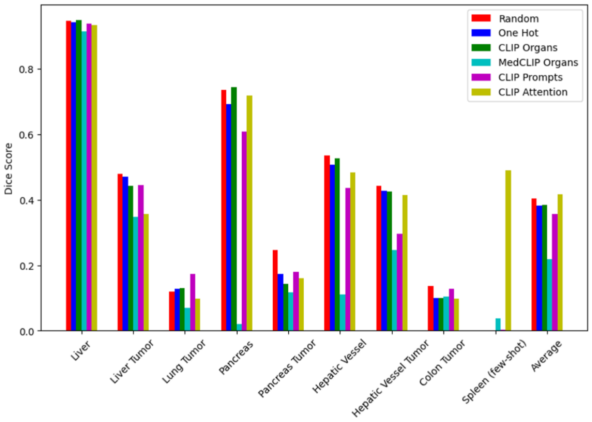

Figure 5: Results

5 Limitation

One potential limitation of our experiments could be the size of the datasets used. We only fine-tuned the original BTCV [1] pretrained model on a portion of the MSD dataset[1]. This means that the results we obtained may not be representative of how the model performs on other, larger datasets. Additionally, another limitation is the computing resources we had available. Our experiments were limited by the amount of processing power and memory we had access to, which could have affected the scale and complexity of our experiments. Future work could address these limitations by utilizing larger datasets and more powerful computing resources to ensure that the results obtained are more robust and generalizable.

6 Conclusion

In conclusion, our study demonstrated the potential benefits of adding language grounding to different tasks by incorporating a fusion module at the representation level in

[1] https://www.synapse.org/#!Synapse:syn25829067/wiki/612712

latent space. However, we also identified the convention for different representations of modality stream before the fusion part could bring some domain bias. We found that simply adding language grounding may not necessarily improve the robustness in segmentation tasks, but it could help with localization tasks. For future work, we propose incorporating bounding box as an object feature as a prior for segmentation tasks and adding 3D spatial information embeddings to our model to further leverage the benefits of language grounding for localization. These potential improvements could enhance the performance and generalizability of our model in various applications.

7 Contributions

The main contribution of our model and experiments is exploring the zero-shot and few-shot capacity of the current state-of-the-art Swin-UNETR[2] model by adding language grounding. By working in this research direction, we aim to create more realistic medical 3D segmentation data that can be used for a wide range of applications in the field of medical imaging. Our findings demonstrate the potential of language grounding in improving the accuracy and generalization of the segmentation models. We hope that our model and experiments can attract more attention to this research direction, leading to further advancements in this area and ultimately benefiting the healthcare industry and patients.

References

[1] M. Antonelli, A. Reinke, S. Bakas, et al. The medical segmentation decathlon. *https://arxiv.org/pdf/1902.09063.pdf*, 2022. 1, 2, 3, 4

[2] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. *arXiv preprint arXiv:2201.01266*, 2022. 2, 3, 4, 5

[3] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. *CoRR*, abs/2103.00020, 2021. 1, 4

[5] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. cite arxiv:1505.04597Comment: conditionally accepted at MICCAI 2015. 2

[6] D. Rozenberszki, O. Litany, and A. Dai. Language-grounded indoor 3d semantic segmentation in the wild. 2022. 2

[7] N. W. Varuna Jayasiri. labml.ai annotated paper implementations, 2020. 2

[8] Z. Wang, Z. Wu, D. Agarwal, and J. Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 2, 4

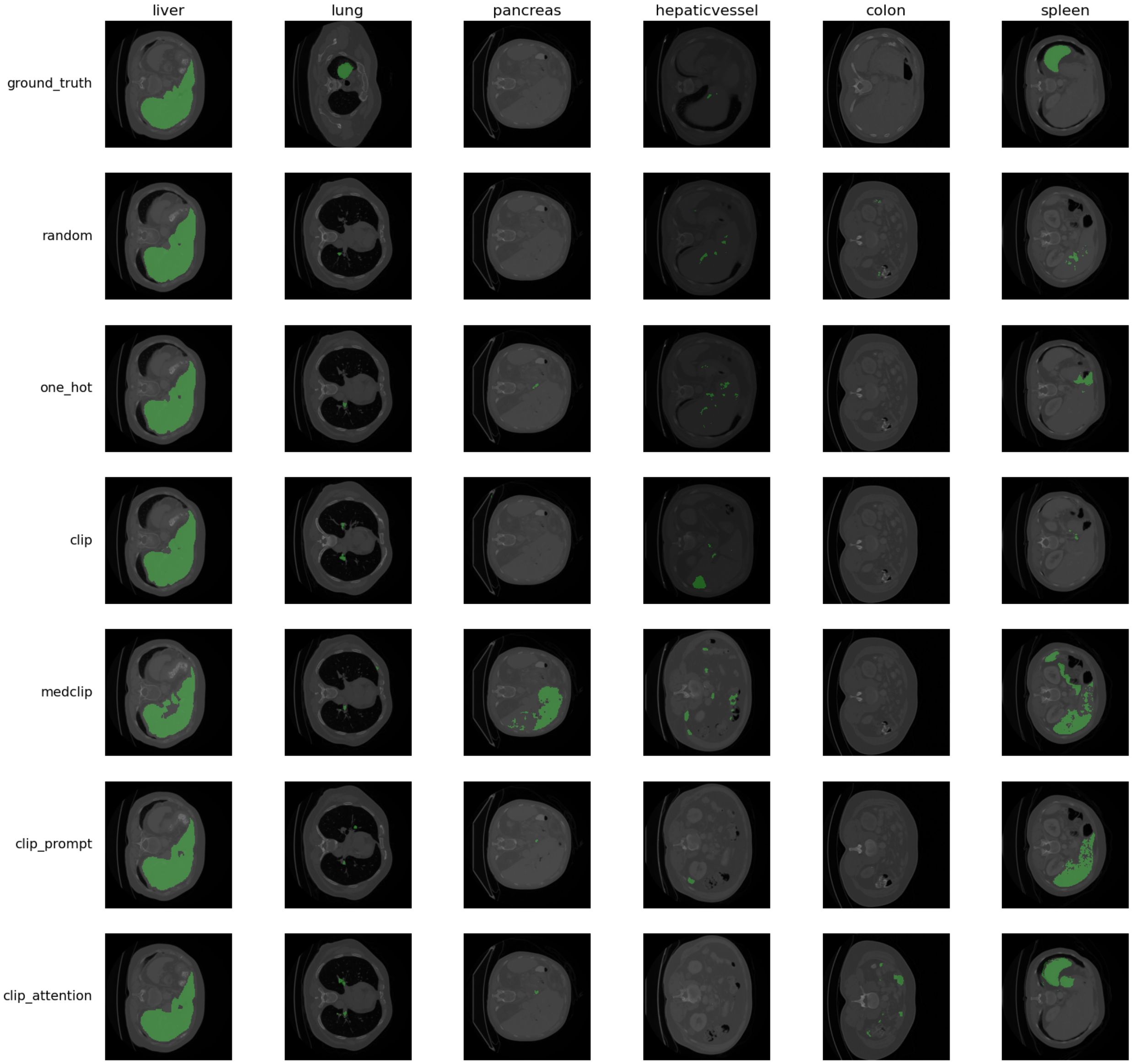

Figure 6: Visual Results1

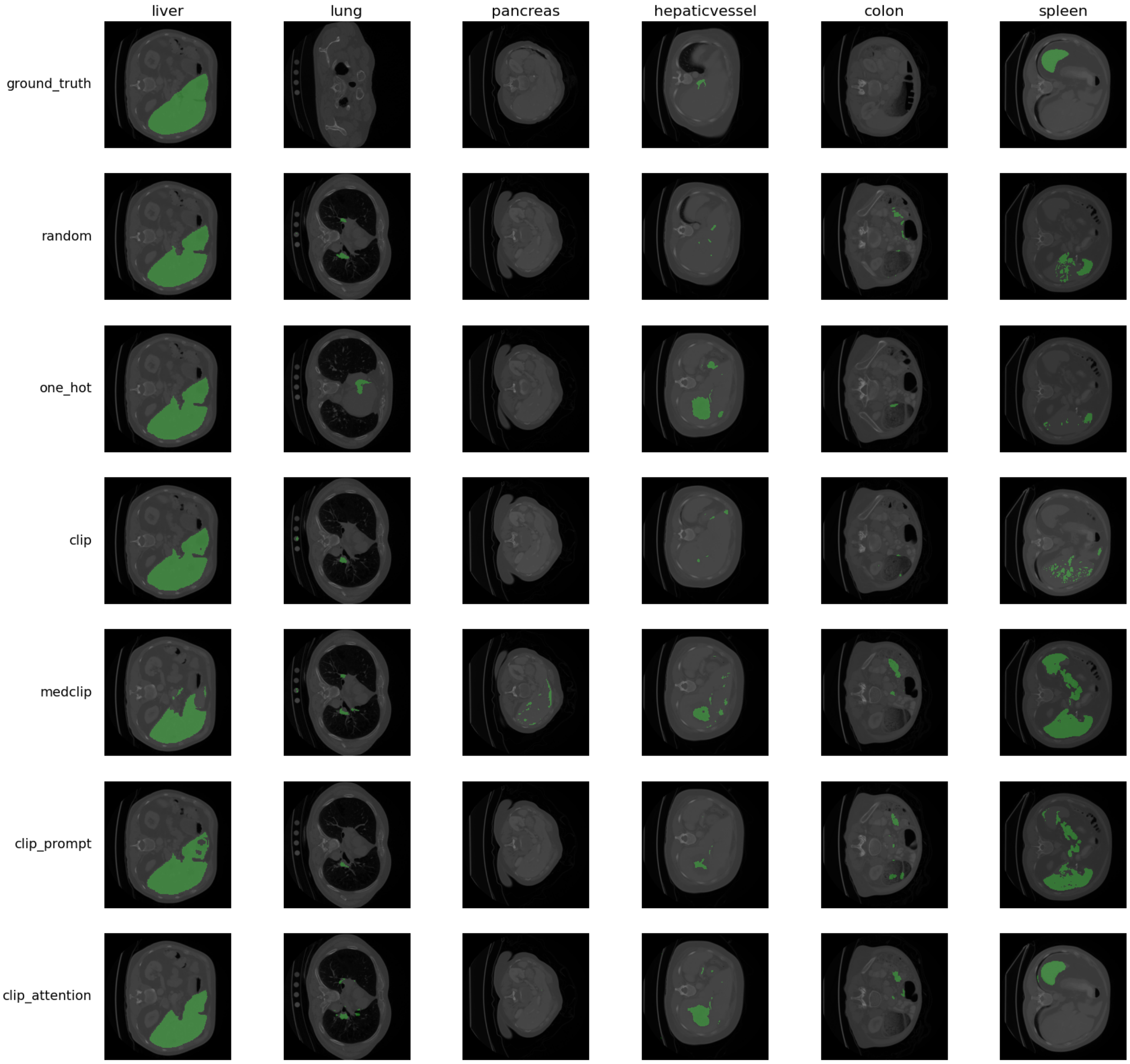

Figure 7: Visual Results2