VLSM-Ensemble: Ensembling CLIP-based Vision-Language Models for Enhanced Medical Image Segmentation

Julia Dietlmeier[1] julia.dietlmeier@insight-centre.org
[1] *Insight Research Ireland Centre for Data Analytics, DCU, Dublin, Ireland*
Oluwabukola Grace Adegboro[*2] oluwabukola.adegboro2@mail.dcu.ie
Vayangi Ganepola[*2] vayangi.ganepola2@mail.dcu.ie
[2] *Research Ireland Centre for Research Training in Machine Learning, DCU, Dublin, Ireland*
Claudia Mazo[3] claudia.mazo@dcu.ie
[3] *School of Computing, Dublin City University, Dublin, Ireland*
Noel E. O'Connor[1] noel.oconnor@insight-centre.org

Editors: Under Review for MIDL 2025

Abstract

Vision-language models and their adaptations to image segmentation tasks present enormous potential for producing highly accurate and interpretable results. However, implementations based on CLIP and BiomedCLIP are still lagging behind more sophisticated architectures such as CRIS. In this work, instead of focusing on text prompt engineering as is the norm, we attempt to narrow this gap by showing how to ensemble vision-language segmentation models (VLSMs) with a low-complexity CNN. By doing so, we achieve a significant performance average Dice gain of 6.3% with the ensembled BiomedCLIPSeg on the BKAI polyp dataset. Furthermore, we provide initial experimental results on the other four radiology and non-radiology datasets. We conclude that ensembling works differently across these datasets (from outperforming to underperforming the CRIS model), indicating a topic for future investigation by the community. The code will be released at https://github.com/juliadietlmeier/VLSM-Ensemble.

Keywords: BiomedCLIP, CLIP, CLIPSeg, Image segmentation, Vision-language models.

1. Introduction

Vision-Language Foundation Models (VLFM) such as CLIP (Radford et al., 2021) and BiomedCLIP (Zhang et al., 2024) are inherently multimodal and are pretrained on large figure-caption datasets. Historically, CLIP was one of the first VLFMs that was pretrained on 400 million image-text pairs from the Internet using contrastive learning. BiomedCLIP is a very recent VLFM specifically created to process biomedical data and is pretrained on 15 million image-text pairs extracted from biomedical research articles in PubMed Central. BiomedCLIP and CLIP can be extended to work in image segmentation scenarios, and some corresponding models are BiomedCLIPSeg (Poudel et al., 2024) and CLIPSeg (Lüddecke and Ecker, 2022). These models consist of separate CLIP-based image and text encoders and a combined image-text decoder. Both CLIP encoders and the decoder have visual transformer-based architectures. During fine-tuning, the CLIP encoders remain frozen and only decoder weights are updated.

[*] Contributed equally

We are inspired by the results frequently achieved in many coding challenges where the winning approaches successfully use ensembling in one form or another (Ferreira et al., 2023). Specifically, we apply one of the ensembling approaches termed *stacking* - a machine learning technique for creating strong models from multiple weak learners.

Our contributions are as follows. In this work, we do not address zero-shot capabilities of VLSMs, but rather focus on joint fine-tuning. We design three ensemble architectures where we combine two VLSMs with a low-complexity UNet-like model: BiomedCLIPSeg with UNet (**BiomedCLIPSeg-A**) and CLIPSeg with UNet (**CLIPSeg-B**). The third architecture is the **Ensemble-C** of BiomedCLIPSeg (Poudel et al., 2024), CLIPSeg (Lüddecke and Ecker, 2022) and UNet (Ronneberger et al., 2015) as shown in Figure 1.

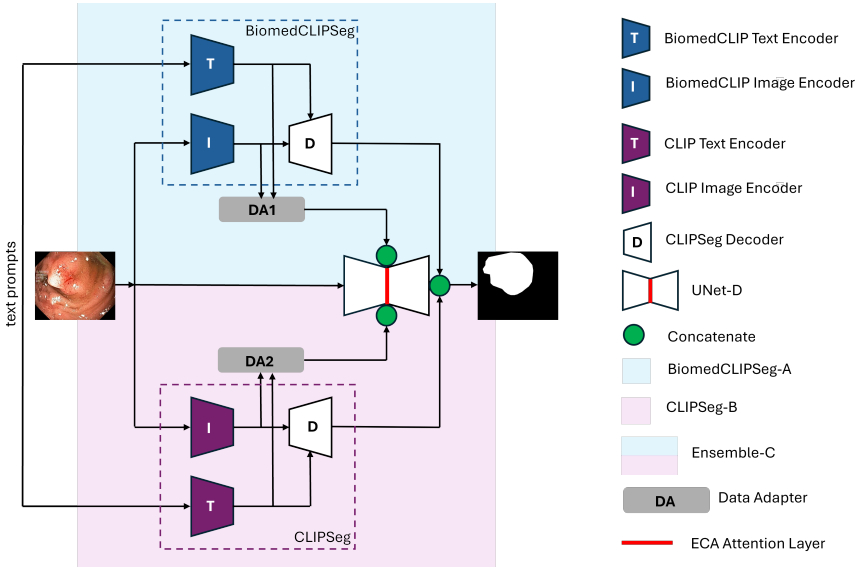

Figure 1: Proposed ensembles of CLIP-based VLSM architectures with the UNet-D.

2. Motivation, Methodology and Results

While designing the ensembling pipelines, the main motivation is to leverage the semantically-rich outputs from the text and image encoders of both BiomedCLIP and CLIP and concatenate them with the bottleneck of a basic UNet model with four encoder-decoder blocks (without a pretrained backbone) and skips. For this purpose, we create two data adapters DA1 and DA2, that adapt each VLSM to the tensor shape of UNet's bottleneck layer. We also integrate an Efficient Channel Attention (ECA) layer (Wang et al., 2020).

Our experimental test bench, text prompts, and data splits are based on the work by (Poudel et al., 2024). We design our ensembling models and prepare the data in such a way that they fit into this framework. We implement all models in PyTorch and train all models (except UNet-D) on our system with Nvidia GeForce RTX 2080 Ti GPU with 11GB VRAM. The UNet-D model was trained on Kaggle using the Tesla P100-PCIE-16GB GPU. We finetune with early stopping with patience 20. We use the AdamW optimizer with an initial learning rate of 10^{-5} and a combination of binary cross-entropy and Dice losses.

The batch size is 10. We evaluated our models on five public datasets: Kvasir-SEG (Jha et al., 2020), ClinicDB (Bernal et al., 2015), BKAI (Lan et al., 2021; An et al., 2022), BUSI (Al-Dhabyani et al., 2020), and CheXlocalize (Saporta et al., 2022).

Table 1: Dice ↑ score results *averaged over all text prompts. Performance gains Δ are shown in gray cell colors such as $\Delta_{CLIPSeg-B}$ is computed relative to CLIPSeg-B.

Model	**Kvasir**[1]	**ClinicDB**[1]	**BKAI**[1]	**BUSI**	**CheXlocalize**
*CRIS (Wang et al., 2022)	0.83043	0.79390	0.85122	0.65453	0.55425
*BiomedCLIPSeg	0.81457	0.75967	0.75120	0.62548	**0.56881**
*BiomedCLIPSeg-A (ours)	0.83976	0.80635	0.81423	0.64908	0.56634
$\Delta_{BiomedCLIPSeg}$	2.519%	4.668%	6.303%	2.36%	-0.247%
*CLIPSeg	0.85744	0.77343	0.82648	0.62271	0.54875
*CLIPSeg-B (ours)	0.86523	0.78966	**0.85234**	0.63529	0.54901
$\Delta_{CLIPSeg}$	0.779%	1.623%	2.586%	1.258%	0.026%
*Ensemble-C (ours)	**0.86795**	**0.81166**	0.84063	**0.66008**	0.52786
$\Delta_{CLIPSeg-B}$	0.272%	2.2%	-1.171%	2.479%	-2.115%
UNet-D (ours)	0.60215	0.33580	0.48360	0.56078	0.29179

Table 1 shows that individual ensembling significantly improves the performance of the BiomedCLIPSeg and CLIPSeg models, especially for the polyp[1] datasets. It can also be seen that the UNet-D component on its own performs poorly across all datasets.

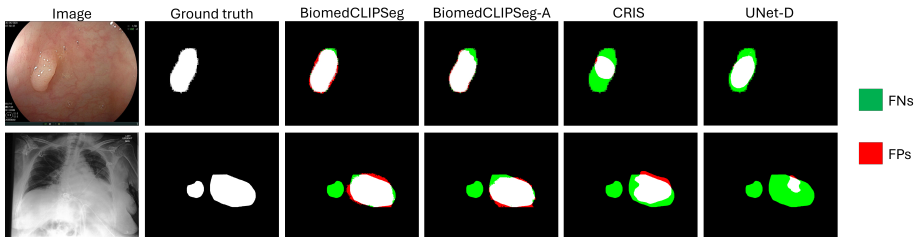

Figure 2: Selected qualitative results for the BKAI (top) and CheXlocalize (bottom).

3. Conclusion and Future Directions

We achieved high average Dice gains on polyp[1] datasets with our ensembled BiomedCLIPSeg-A model. The selected qualitative results in Figure 2 show fewer False Positives (FPs) for the BiomedCLIPSeg-A compared to BiomedCLIPSeg and fewer False Negatives (FNs) compared to CRIS. However, neither the VLSM model recovers the small mask on the left, while the CRIS model has the most FNs. In the future, we anticipate understanding the reason for some negative gains, the global and individual relationships between quantitative and qualitative results, and investigating the impact of text prompts on performance.

Acknowledgments

This work was funded by the Insight Research Ireland Centre for Data Analytics and partly supported by Taighde Éireann – Research Ireland under Grant number 18/CRT/6183 (Adegboro, Ganepola).

References

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

Nguyen S An, Phan N Lan, Dao V Hang, Dao V Long, Tran Q Trung, Nguyen T Thuy, and Dinh V Sang. Blazeneo: Blazing fast polyp segmentation and neoplasm detection. *IEEE Access*, 10:43669–43684, 2022.

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.

André Ferreira, Naida Solak, Jianning Li, Philipp Dammann, Jens Kleesiek, Victor Alves, and Jan Egger. How we won BraTS 2023 adult glioma challenge? just faking it! enhanced synthetic data augmentation and model ensemble for brain tumour segmentation. *MICCAI*, 2023.

Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *International Conference on Multimedia Modeling*, pages 451–462. Springer, 2020.

Phan Ngoc Lan, Nguyen Sy An, Dao Viet Hang, Dao Van Long, Tran Quang Trung, Nguyen Thi Thuy, and Dinh Viet Sang. Neounet: Towards accurate colon polyp segmentation and neoplasm detection. *In Advances in Visual Computing: 16th International Symposium, ISVC*, Part II:15–28, 2021.

Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. *CVPR*, 2022.

Kanchan Poudel, Manish Dhakal, Prasiddha Bhandari, Rabin Adhikari, Safal Thapaliya, and Bishesh Khanal. Exploring transfer learning in medical image segmentation using vision-language models. *MIDL*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. *MICCAI*, 2015.

Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, and Andrew Y et al. Ng. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022.

Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient channel attention for deep convolutional neural networks. *CVPR*, 2020.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. CRIS: CLIP-driven referring image segmentation. *CVPR*, 2022.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Andrea Tupini, Yu Wang, Matt Mazzola, Swadheen Shukla, Lars Liden, Jianfeng Gao, Angela Crabtree, Brian Piening, Carlo Bifulco, Matthew P. Lungren, Tristan Naumann, Sheng Wang, and Hoifung Poon. A multimodal biomedical foundation model trained from fifteen million image–text pairs. *NEJM AI*, 2(1), 2024. doi: 10.1056/AIoa2400640. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400640.