MSc Thesis

Med-VLM: Prompt Adaptation of Vision-Language Models to Medical Domains

Abstract

Medical imaging faces persistent domain shifts across scanners, protocols, pathologies, and populations. Recent vision-language models, such as CLIP [4] (Contrastive Language-Image Pre-training), learns joint representations from image-text pairs, enabling zero-shot transfer to new visual concepts through natural language descriptions. Although the recent works demonstrate that CLIP [4] based models offer strong zero-shot transfer, their performance degrades on unseen medical data. Recent adaptation techniques aim to mitigate this by proposing prompt-based, parameter-efficient, and feature-level strategies [1, 2]. Complementing this, text-only prompt learning [3, 5, 6] optimizes prompts in the text space using LLM-derived biomedical descriptions or ontology-based prototypes, improving transfer to unseen data. This Master's thesis proposes adapting CLIP-based models to unseen medical data through techniques including prompt learning and CLIP's consistent contrastive alignment, emphasizing practical clinical feasibility.

This Master thesis aims to analyze and propose: (i) an adaptation framework that leverages medical concepts into prompts to improve robustness on unseen data, (ii) prompt-based adaptation procedures for non-independent and identically distributed (non-i.i.d) shifts through lightweight mechanisms, (iii) a comprehensive evaluation on multi-source medical benchmarks [2]. Expected outcomes include, but are not limited to: a vision and language-guided framework for medical imaging that improves generalization across multiple scanners and diverse population data, a lightweight test-time adaptation method with biomedical prompts, and analysis on medical benchmarks with ablations under realistic scenarios. This Master's thesis aspires to publish results in a relevant academic venue.

Requirements

Experience with practical deep learning frameworks such as PyTorch, and interest in medical imaging research. Interest in the fundamentals of deep learning techniques would be an added advantage.

Application

Please send an email, involving a CV, a current transcript of records, and a brief statement on why you are interested in the project, to sameer.ambekar@tum.de.

Affiliation

Prof. Dr. Julia Schnabel
Informatik 32 - Lehrstuhl für Computational Imaging and AI in Medicine

Supervision: Sameer Ambekar, Dr. Daniel M. Lang

References

[1] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao. Biomedcoop: Learning to prompt for biomedical vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14766–14776, 2025.

[2] X. Li, J. Li, F. Li, L. Zhu, Y. Yang, and H. T. Shen. Generalizing vision-language models to novel domains: A comprehensive survey. *arXiv preprint arXiv:2506.18504*, 2025.

[3] S. Pratt, U. Jain, et al. Cupl: Prompting clip with (free) language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Chen, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[5] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[6] K. Zhou, J. Yang, C. C. Loy, and Z. Liu. Learning to prompt for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.