Project Description

Reliable Vision-Language Models for Healthcare Applications

Abstract

Recent advances in multimodal deep learning have the potential to achieve unprecedented levels of automation in addition to high levels of accuracy in tasks such as visual question answering, phrase grounding, caption generation, and image retrieval based on text. In particular, advances that combine large language models (LLMs) with vision encoders bring unprecedented capabilities that are leverageable under a wide range of healthcare applications. Nevertheless, the reliability of these technologies has yet to be comprehensively examined, especially in crucial domains such as the aforementioned healthcare, where erroneous decisions may lead to extremely severe consequences.

The proposed research aims to address this critical gap by investigating methods designed to quantify the uncertainty inherent in such multimodal systems. Our primary focus will be on vision-language models (VLM), which are increasingly finding applications across a wide range of domains. The proposed methodology will be utilised to develop VLM-based frameworks with a strong emphasis on reliability, subsequently benchmarking them using biomedical data.

Introduction

The importance of automating routine workflows in healthcare to enhance the capacity for combating diseases cannot be overstated. By automating these processes, we can liberate clinicians' time, allowing them to focus on more intellectually demanding tasks. Multimodal deep learning possesses the potential to address this issue significantly, as evidenced by previous work [1]. Integration of various information modalities, such as health records, X-ray images, and MRI scans, brings the capacity to provide comprehensive responses to a wide array of inquiries, all without requiring human intervention [2]. Nonetheless, the practical adoption of these systems remains limited, primarily due to the inherent risks associated with such models. These systems, at times, have the potential to yield incorrect results with relatively high confidence, as evidenced by prior research [3,4,5]. Consequently, there is a pressing need for the establishment of a framework that incorporates measures of *reliability* and *uncertainty*, ensuring timely intervention whenever necessary.

Research Challenge

The most prevalent form of multimodal deep models is the Vision-Language Model (VLM) [1,6,7]. These models are guided not only by visual features but also by corresponding textual information. However, the research landscape pertaining to uncertainty estimation in VLMs is nascent and somewhat constrained. Within this field, two primary approaches have emerged.

The first assumes a static VLM and aims to gauge uncertainty specifically for downstream tasks, as illustrated in earlier study [8]. In contrast, the second approach modifies a VLM by adapting the training procedures of these models, primarily through adjustments to the loss function, with the aim of integrating the concept of reliability into the training process [9, 10].

Both methodologies have demonstrated significant improvements over baseline models, particularly in terms of reliability, and hold promise for applications in phenotyping and treatment selection. However, it is imperative to critically evaluate and acknowledge several key considerations that affect their practical utility:

1. **Task-Specific Uncertainty vs. Generalization**: The first method does not necessitate retraining of the VLM, which can be computationally demanding. However, its uncertainty

estimates are intricately linked to the specific downstream task, necessitating adaptation for each new task. This adaptability may prove impractical for tasks with limited data, ultimately limiting the zero-shot capabilities of the VLM.

2. **Decomposition of Uncertainty Across Modalities**: The first class of methods assumes a decomposition of uncertainty across different modalities. While this approach simplifies the modeling process, it may fall short in capturing the complex interrelationships between different modalities, potentially undermining the model's ability to capture cross-modal dependencies effectively.
3. **Reliance on Point Estimates**: A critical concern with the first type of methods is their heavy reliance on the quality of point estimates generated by the VLM for computing uncertainty estimates. This reliance can be problematic when data distributions change, such as with the introduction of new diseases, drugs, or diagnostic methods, which may render the point estimates unreliable. This introduces a vulnerability in the model's generalizability to evolving data.
4. **Arbitrary Nature of Loss Modifications**: The second class of methods, which involves the modification of the loss function to incorporate a reliability component, raises concerns regarding the arbitrary nature of this connection. The theoretical underpinning of this approach is often lacking, and the link between the added loss component and the actual uncertainty estimates may be tenuous. Consequently, while these methods demonstrate consistent performance across domains, the reliability of their uncertainty estimates remains a subject of concern, as they might not be sufficiently accurate.

Addressing these research challenges is pivotal for the practical and robust implementation of VLMs in healthcare and other critical domains.

Data & Methodology

The domains of vision and language have seen significant developments towards probabilistic modelling, uncertainty estimation, and reliability. We propose to leverage the relevant work done in these unimodal settings. For instance, [11] uses probabilistic objectness to efficiently identify previously unknown objects for object detection tasks, an endeavour that has posed considerable challenges when approached outside the probabilistic framework. Even within unimodal healthcare applications, such as brain tumour segmentation, the adoption of probabilistic frameworks has yielded tangible benefits [12]. Moreover, the field of deep probabilistic language models has been advancing at a rapid pace. [13] focuses on incorporating latent stochastic processes into language modelling, resulting in improved performance and reliability across numerous benchmarking tasks. Intriguingly, [14] contends that large language models, trained on extensive datasets, inherently exhibit robust probabilistic properties. They even put forth gradient-based sampling methods that effectively estimate the uncertainty associated with these large language models.

In most cases, multimodal training procedures involve the utilization of techniques like CLIP [15], BLIP [16], BLIP2 [17] contrastive alignment-based fine-tuning techniques, which can modify the vision encoder or the language model. These modifications can, in turn, alter the nature of the uncertainty estimates provided by these models, often rendering them seemingly less useful.

We conjecture that unimodal deep probabilistic models retain their fundamental characteristics and simply adapt to the new objective during the fine-tuning process, such as in the cases of CLIP or BLIP(2). In light of this hypothesis, a comprehensive exploration of the nature of these adaptations, both from a theoretical and empirical perspective will be undertaken. Further, a robust framework for interpreting these adaptations, within commonly employed objectives, will be established. These steps will enable the utilisation of the extensive available literature for unimodal applications within multimodal contexts. Both the vision and language models will become more readily replaceable and adaptable, as is often the case when a point estimate is the primary requirement.

This research direction will unveil novel possibilities for designing fine-tuning objectives that explicitly retain the probabilistic essence of the underlying models, consequently enabling accurate uncertainty estimates in the cross-domain context of image and text. Pragmatic use of these systems will be evidenced by considering tasks such as guided MRI scan transfer, chest X-ray analysis, and MRI analysis using publicly available data.

RRI / Ethical Considerations

We identify several ethical and RRI considerations that require to be carefully addressed throughout the project lifecycle. *Ethical Considerations:* The project will undertake research activities in medical imaging and healthcare related domains. Ethical considerations such as maintaining integrity in data collection and analysis, must be integrated into research protocols.This includes adherence to ethical guidelines and obtaining ethical approvals where necessary for research involving sensitive data. *Responsible Innovation:* we endeavour to adopt inclusive methodologies in all stages of project research and innovation, from agenda setting to design, implementation, and evaluation. This approach will ensure that the research outcomes are socially and ethically acceptable, towards addressing real-world challenges effectively. *Open Access and Data Sharing:* the project will promote open access to research findings and data to maximise transparency, collaboration and knowledge exchange. This involves adopting open access publication practices as standard practice, sharing research data openly where possible, and adhering to data sharing protocols while respecting intellectual property rights and any confidentiality agreements with industry, government and third-party stakeholders.

Expected Outcome & Impact

Findings of the research will be disseminated in top-tier peer-reviewed conferences and journals such as ICCV, ECCV, MLHC, NeurIPS, ICML, TMI, PAMI, and JMLR. Furthermore, we commit to making the code for our research publicly available on GitHub, thereby facilitating future research and promoting transparency and collaboration in the scientific community. Moreover, the university's commitment to fostering interdisciplinary collaborations and training will be taken advantage of to help amplify impact. Initiatives like Healthcare AI, Computational Biology, and Bioinformatics provide an ideal environment for knowledge exchange and partnership opportunities.

References

1. Huang, S., Shen, L., Lungren, M. P., & Yeung, S. (2021). GLORIA: a multimodal Global-Local Representation learning framework for label-efficient medical image recognition. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. https://doi.org/10.1109/iccv48922.2021.00391
2. Zhang, C., Chu, X., Ma, L., Zhu, Y., Wang, Y., Wang, J., & Zhao, J. (2022). M3Care: Learning with Missing Modalities in Multimodal Healthcare Data. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. https://doi.org/10.1145/3534678.3539388
3. Whitehead, S., Petryk, S., Shakib, V., Gonzalez, J. E., Darrell, T., Rohrbach, A., & Rohrbach, M. (2022). Reliable visual question answering: abstain rather than answer incorrectly. In Lecture Notes in Computer Science (pp. 148–166). https://doi.org/10.1007/978-3-031-20059-5_9
4. McKnight, D. H., Carter, M., Thatcher, J. B., & Clay, P. (n.d.). Trust in a specific technology. ACM Transactions on Management Information Systems, 2(2), 1–25. https://doi.org/10.1145/1985347.1985353

5. Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: Focus on clinicians. Journal of Medical Internet Research, 22(6), e15154. https://doi.org/10.2196/15154
6. Zhang, Y., Merck, D., Tsai, E. B., Manning, C. D., & Langlotz, C. P. (2020). Optimizing the Factual Correctness of a Summary: A Study of Summarizing Radiology Reports. ACL Anthology. https://doi.org/10.18653/v1/2020.acl-main.458
7. Miura, Y., Zhang, Y., Tsai, E. B., Langlotz, C. P., & Jurafsky, D. (2021). Improving Factual Completeness and Consistency of Image-to-Text Radiology Report Generation. ACL Anthology. https://doi.org/10.18653/v1/2021.naacl-main.416
8. Upadhyay, U., Karthik, S., Mancini, M., & Akata, Z. (2023). ProbVLM: Probabilistic Adapter for Frozen Vison-Language Models. https://openaccess.thecvf.com/content/ICCV2023/html/Upadhyay_ProbVLM_Probabilistic_Adapter_for_Frozen_Vison-Language_Models_ICCV_2023_paper.html
9. Chun, S., Oh, S. J., De Rezende, R. S., Kalantidis, Y., & Larlus, D. (2021). Probabilistic Embeddings for Cross-Modal Retrieval. CVPR. https://doi.org/10.1109/cvpr46437.2021.00831
10. Song, Y., & Soleymani, M. (2019). Polysemous Visual-Semantic Embedding for Cross-Modal Retrieval. CVPR. https://doi.org/10.1109/cvpr.2019.00208
11. Zohar, O., Wang, K., & Yeung, S. (2023). PROB: Probabilistic Objectness for Open World Object Detection. CVPR. https://doi.org/10.1109/cvpr52729.2023.01101
12. Kamnitsas, K., Bai, W., Ferrante, E., McDonagh, S., Sinclair, M., Pawlowski, N., Rajchl, M., Lee, M. C. H., Kainz, B., Rueckert, D., & Glocker, B. (2018). Ensembles of multiple models and architectures for robust brain tumour segmentation. In Lecture Notes in Computer Science (pp. 450–462). https://doi.org/10.1007/978-3-319-75238-9_38
13. Wang, R. E., Durmus, E., Goodman, N., & Hashimoto, T. (2022). Language modeling via stochastic processes. International Conference on Learning Representations. Retrieved from https://openreview.net/forum?id=pMQwKL1yctf
14. Kumar, S., Paria, B., & Tsvetkov, Y. (2022). Gradient-based Constrained Sampling from Language Models. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.144
15. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *International Conference on Machine Learning*.
16. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2201.12086
17. Li, J., Li, D., Savarese, S., & Hoi, S. C. H. (2023). BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2301.12597