Stress-Testing of Multimodal Models in Medical Image-Based Report Generation

Flávia Carvalhido, Henrique Lopes Cardoso, Vítor Cerqueira

LIACC - Artificial Intelligence and Computer Science Laboratory
Faculty of Engineering, University of Porto
Rua Dr. Roberto Frias
4200-465 Porto, PORTUGAL
up201806857@up.pt

Abstract

Multimodal models, namely vision-language models, present unique possibilities through the seamless integration of different information mediums for data generation. These models mostly act as a black-box, making them lack transparency and explicability. Reliable results require accountable and trustworthy Artificial Intelligence (AI), namely when in use for critical tasks, such as the automatic generation of medical imaging reports for healthcare diagnosis. By exploring stress-testing techniques, multimodal generative models can become more transparent by disclosing their shortcomings, further supporting their responsible usage in the medical field.

Introduction

The recent research boom on Large Language Models (LLM) and generative AI has made it simple to streamline data generation. By using millions of parameters, these models can effectively achieve impressive performance across a variety of benchmark tasks (Wei et al. 2022).

Context

While generation capabilities are commonly associated with LLM, other generative models have been developed leveraging several mediums of information. Notably, image-text multimodal models showcase sizeable strides in development, using LLMs' excelling abilities for generation based on visual information.

The excellent encoding capabilities these models present make them specially useful for tasks that involve both types of data – image and text. Their application in medical image report generation has been explored (Shamshad et al. 2023), as these methodologies can aid in radiology exam diagnosis, making it a faster process and less prone to human error.

Although promising, such large neural models have their pitfalls: requiring too much data and computing resources during training, replicating biased or wrong information during generation, hallucinations, lack of reasoning capabilities, and others (Wei et al. 2022; Ji et al. 2023). From their training data to their black-box architectures, there are a multitude of characteristics that undermine the transparency and reliability of generative AI models.

Motivation

In tasks such as the generation of medical imaging reports, having transparent and explainable AI is imperative. The lack of trustworthiness in AI is detrimental to its own usage, as an incorrectly generated report could translate into serious consequences to patients' health. It is essential to understand the model's interpretation of a medical image and reasoning behind a diagnostic.

Fostering the ethical development and use of AI models is a vital step for attaining trustworthy AI. Unveiling details on a model's training process is essential for its informed usage. Even though responsible AI is a pressing issue and various AI regulations (Siau and Wang 2020) have been created, there is a lack of practical approaches for guaranteeing responsible AI development.

Stress-testing is a common process for software development aimed at understanding the shortcomings of software programs by studying their behavior under different scenarios. While it is not a standard procedure to stress-test AI models, due to their complex inner-workings which difficult interpretability, it could be the answer towards identifying some limitations of these technologies (Naik et al. 2018), namely within the task of medical image report generation.

Hypothesis and Research Questions

This thesis' hypothesis is as follows: *It is possible to enhance the responsible usage of generative multimodal models for medical image report generation by extensively identifying their shortcomings through stress-testing.*

Furthermore, the research problem is formulated within four research questions:

- **RQ1**: Which types of limitations should be looked for in image-text multimodal models, taking into account the need for explainable AI?
- **RQ2**: What stress-testing approaches are best-fit for multimodal models?
- **RQ3**: Which are the limitations of image-text multimodal models, both generally and in the task of medical imaging report generation?
- **RQ4**: How much can identifying limitations enhance the explainability and responsible use of multimodal models?

Task Name	Start Date	End Date
Initial Literature Review	Jan 2024	Sep 2024
Multimodal Dataset Select.	July 2024	Oct 2024
Multimodal Models Select.	Sep 2024	Nov 2024
Stress-testing Methods Def.	Oct 2024	Jan 2025
Research Pipeline Preparation	Nov 2024	Feb 2025
Multimodal Models Training	Jan 2025	Jan 2026
Stress-testing Models	Mar 2025	Mar 2026
Result Gathering & Discuss.	Jan 2026	Sep 2026
Final Document Writing	Jun 2026	Mar 2027

Table 1: Planed research timeline.

Related Work

Several Vision-Language Models (VLM) have been proposed for medical applications (Shamshad et al. 2023). Leveraging an encoder-decoder structure and the Transfomer architecture, methodologies such as AlignTransformer (You et al. 2021), MMTN (Cao et al. 2023), and RGRG (Tanida et al. 2023) are highlighted for their performance on medical image report generation.

Many authors have discussed the issue of AI explainability in healthcare (Morley et al. 2020), further classifying it into different ethical issues. Some practical methods have been proposed within the Explainable AI field to address said issues and provide more clarity into AI decision-making in the medical field (Yang et al. 2023). However, most incur in a prediction vs. explainability trade-off and fail to pinpoint the general limitations of the models.

Planned Timeline

As this work started in January 2024, with a planning phase which lasted until March 2024, a timeline was prepared further dividing the main tasks and assessing a time frame for each activity, as shown in Table 1.

It is worth noting that the planned work is to be carried out by me under the supervision of Professors Henrique Lopes Cardoso and Vítor Cerqueira.

Until September 30th 2024, a thorough literature review has been conducted. The dataset and model selection is underway. Some initial difficulties concerning the proposal of stress-testing methodologies were felt, due to the multitude of limitations which can be studied, which we are defining as part of RQ1 to better focus our research efforts.

As of February 25th 2025, we plan to have a deeper understanding of the selected methodologies and devise stress-testing procedures accordingly. Given that the stress-testing methodologies definition is a crucial step of this work, we believe the AAAI'25 Doctoral Symposium comes at a vital moment in the planned work timeline, as it will allow for discussion of our proposed stress-testing strategies.

Acknowledgements

Work funded by Agenda "Center for Responsible AI", nr. C645008882-00000055, investment project nr 62, financed by the Recovery and Resilience Plan (PRR) and by European Union - NextGeneration EU; and, Base Funding - UIDB/00027/2020 and Programatic Funding - UIDP/00027/2020 of the Artificial Intelligence and Computer Science Laboratory (LIACC) funded by national funds through the FCT/MCTES (PIDDAC).

References

Cao, Y.; Cui, L.; Zhang, L.; Yu, F.; Li, Z.; and Xu, Y. 2023. MMTN: multi-modal memory transformer network for image-report consistent medical report generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 277–285.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12): 1–38.

Morley, J.; Machado, C. C.; Burr, C.; Cowls, J.; Joshi, I.; Taddeo, M.; and Floridi, L. 2020. The ethics of AI in health care: a mapping review. *Social Science & Medicine*, 260: 113172.

Naik, A.; Ravichander, A.; Sadeh, N.; Rose, C.; and Neubig, G. 2018. Stress Test Evaluation for Natural Language Inference. In Bender, E. M.; Derczynski, L.; and Isabelle, P., eds., *Proceedings of the 27th International Conference on Computational Linguistics*, 2340–2353. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Shamshad, F.; Khan, S.; Zamir, S. W.; Khan, M. H.; Hayat, M.; Khan, F. S.; and Fu, H. 2023. Transformers in medical imaging: A survey. *Medical Image Analysis*, 88: 102802.

Siau, K.; and Wang, W. 2020. Artificial intelligence (AI) ethics: ethics of AI and ethical AI. *Journal of Database Management (JDM)*, 31(2): 74–87.

Tanida, T.; Müller, P.; Kaissis, G.; and Rueckert, D. 2023. Interactive and Explainable Region-guided Radiology Report Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682.

Yang, W.; Wei, Y.; Wei, H.; Chen, Y.; Huang, G.; Li, X.; Li, R.; Yao, N.; Wang, X.; Gu, X.; et al. 2023. Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3(3): 161–188.

You, D.; Liu, F.; Ge, S.; Xie, X.; Zhang, J.; and Wu, X. 2021. Aligntransformer: Hierarchical alignment of visual regions and disease tags for medical report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24*, 72–82. Springer.