Improving Reliability in Vision-Language Models: Reducing Hallucination in Multimodal AI Systems

Introduction

Recently, large vision-language models (LVLMs) have made significant strides, demonstrating impressive capabilities in multimodal interaction. However, these models often suffer from hallucination, where generated textual outputs fail to align with the visual input. For example, models may describe non-existent objects or provide fabricated context (as shown in Fig. 1), undermining their reliability, particularly in high-stakes domains such as healthcare and autonomous driving systems.

This doctoral project aims to investigate the fundamental factors contributing to hallucination, develop innovative strategies to mitigate it and establish new benchmarks and metrics for evaluation. This work is to contribute to the development of the next generation of reliable and ethically sound LVLMs, advancing both academic theory and practical applications.

Fig. 1 Some typical hallucination examples in LVLMs [1].

Research Questions

- What are the primary causes of hallucination in LVLMs?
- How can datasets, model architectures, and training processes be improved to reduce hallucination?
- What metrics and benchmarks are most effective in measuring and mitigating hallucination?
- How can these advancements enhance the availability of LVLMs in real-world applications?

Methodology and Timeliness

Year 1: Diagnosing Hallucination

- **Dataset Analysis:** Examine widely used datasets (e.g., MSCOCO) for biases and noise analysis.
- **Model Analysis:** Conduct experiments and ablation studies on leading LVLMs to identify hallucination-inducing factors, such as multimodal fusion strategies and deviations in attention mechanisms.
- **Taxonomy Development:** Develop a taxonomy which classifies hallucination types (e.g., context hallucination, syntax hallucination).

Year 2: Developing Mitigation Strategies

- **Dataset Refinement:** Create denoising modules to enhance visual-textual grounding.
- **Architectural Innovations:** Test improved models such as attention mechanisms and constrained decoding methods.

Year 3: Evaluation and Benchmarking

- **Metrics Development:** Define hallucination-specific metrics, such as grounding fidelity and factual accuracy.
- **Benchmarking:** Propose new benchmarks alongside existing datasets for robust evaluation.
- **Human-in-the-Loop Testing:** Validate the improvements of models in practical, real-world scenarios.

Year 4: Synthesis and Dissemination

- Consolidate findings into a comprehensive framework for reducing hallucination in LVLMs.
- Publish findings in top-tier conferences and journals and write up thesis.
- Disseminate datasets, benchmarks, and open-source tools to promote academic and industrial collaboration.

Impact and Relevance

This research tackles a critical challenge in multimodal AI (a cornerstone of realising artificial general intelligence), enhancing the reliability and usability of LVLMs. By reducing hallucination, the project will enable their deployment in high-stakes fields, such as healthcare and accessibility, where factual grounding is essential. Academic contributions include the development of new metrics, benchmarks, and algorithms, while practical outcomes involve collaboration with industries to apply these findings. Public engagement efforts, such as workshops and open-source tools, will ensure widespread dissemination and knowledge exchange.

Key Skills Through This Project

Successful candidate has an opportunity to work in an international research group in Microsystem Group, Newcastle University. The key skills the candidate can acquire are as follows:

Research and Analytical Skills

- Critical Analysis of Model Behaviour
    - Investigating model limitations and identifying factors contributing to hallucination.
    - Performing error analysis and ablation studies.
- Experiment Design and Hypothesis Testing
    - Designing robust experiments to test new mitigation strategies.
    - Statistical analysis of results.
- Programming Languages and Frameworks
    - Proficiency in Python and libraries like PyTorch or TensorFlow.
    - Experience with libraries such OpenCV, and torchvision.

Other Skills for Future Career

- Scientific Communication
    - Writing research papers, technical reports, and presenting findings at top conferences and journals.
    - Presentation skills to interdisciplinary audiences.
- Collaboration and Teamwork
    - Working in cross-disciplinary teams with diverse backgrounds.
    - Managing collaborative codebases using tools like GitHub.
- Project Management
    - Planning long-term research projects.
    - Time management and setting milestones to achieve research goals.

By acquiring these skills, applicants will position themselves as experts in both the academic and applied aspects of vision-language models, ready to tackle cutting-edge challenges in AI.

Feasibility

This project leverages existing datasets, state-of-the-art LVLMs, and available computational resources from the School of Engineering, Newcastle University. All experiments will adhere to ethical guidelines, including data privacy. By advancing methods to diagnose, mitigate, and evaluate hallucination, this project will set new standards for grounded, trustworthy AI systems, benefiting both academia and industry.

Entry Requirements

Essential:

- 2.1 or MSc degree (or equivalent) in Computer Science, or a closely related subject.

- Proficient programming skills (Python/C/C++/MATLAB.

Desirable:

- Experience of LVLM algorithms and frameworks such as Pytorch.

Contact

Please do contact Dr. Zhuang Shao, email: zhuang.shao@newcastle.ac.uk for further enquiries if you are interested in this opportunity.

References

[1] Liu H, Xue W, Chen Y, et al. A survey on hallucination in large vision-language models[J]. arXiv preprint arXiv:2402.00253, 2024.