Bhathiya Hemanthage

School of Mathematical and Computer
Sciences.
Heriot-Watt University,
Riccarton, Edinburgh

hsb2000@hw.ac.uk

1 Research interests

My research focus on **Visual Dialogues** and **Generalized Visual-Language Grounding with Complex Language Context**. Specifically, my research aim to utilize Large Language Models (LLMs) to build *conversational agents capable of comprehending and responding to visual cues*.

Visual-Language Pre-trained (VLP) models, primarily utilizing transformer-based encoder-decoder architectures, are extensively employed across a range of visual-language tasks, such as visual question answering (VQA) and referring expression comprehension (REC). The effectiveness of these models stems from their robust visual-language integration capabilities. However, their performance is constrained in more complex applications like multimodal conversational agents, where intricate and extensive language contexts pose significant challenges. These tasks demands language-only reasoning before engaging in multimodal fusion. In response, my research investigates the application of Large Language Models (LLMs) with advance comprehension and generation capabilities to enhance performance in complex multimodal tasks, particularly multimodal dialogues.

In brief, my work in visual dialogues revolves around two major research questions. i) How to redefine visually grounded conversational agent architectures to benefit from LLMs ii) How to transfer the large body of knowledge encoded in LLMs to conversational systems.

1.1 End-to-end multi-modal conversational agents with Symbolic Scene Representation

The SIMMC 2.0 (Kottur et al., 2021) is a multi-modal task oriented dialogue proposed as part of DSTC-10 challenge with the goal of facilitating task oriented dialogue system which can understand visual context in addition to the linguistic context. This is challenging compared to both text-only dialogue datasets (such as ()) and image querying dialogue (such as ()) due to the simultaneous presence of signals from multiple modalities, which a user can refer to at any point in the conversation.

Despite the inherent complexity of multimodal dialogues, our work; (Hemanthage et al., 2023) introduce SimpleMTOD, which recasts all sub-tasks into a simple language model. In (Hemanthage et al., 2023) , we represent the visual information in a symbolic manner. SimpleMTOD combines de-localized object representation with token based spatial information representation.

However, (Hemanthage et al., 2023) and most other works on multimodal dialogue systems (Chen et al., 2023; Long et al., 2023) make a key unrealistic assumption in their inference processes. They assume the availability of a predefined catalog of items that may appear in a scene, and that this catalog remains fixed from training to inference. Our current work (Hemanthage et al., 2024) focus on addressing limitations in real-world applicability due to these unrealistic assumptions.

1.2 Modular multi-modal conversational agents with Pseudo-Labelling

End-to-end modeling with multimodal fusion has demonstrated significant advancements in various visual-language tasks, including phrase grounding (Plummer et al., 2015), referring expression comprehension (REC) (Yu et al., 2016; Nagaraja et al., 2016), and open vocabulary object detection (Gu et al., 2021). However, the applicability of these methods is limited when the language context is sophisticated, as in visual dialogues.

Modular approaches presents several advantages for the more complex multi-modal dialogue task. Firstly, modules can be designed to enable explicit text-only reasoning over the dialogue context, which is crucial in visual dialogue systems. Secondly, the modular approaches can mitigate the challenges posed by lengthy language contexts by summarizing the language context to extract only the essential information for the task before visual-language fusion.

Despite the advantages, a key challenge of the modular approach is the lack of annotated data for training intermediate modules. To address this, our work (Hemanthage et al., 2024), explore semi-supervised learning (SSL) setup where pseudo-labels generated by prompting a Large Language Model (LLM) serve as training targets for intermediate modules. Although our work focuses on Ambiguous Candidate Identification (ACI) in multimodal dialogues, the general approach of modularization with LLM-based pseudo-labelling can be extended to other

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 57–59
September 16–17, 2024. ©2024 Association for Computational Linguistics

complex multimodal tasks with long language context.

2 Spoken dialogue system (SDS) research

Considering the remarkable advancements in artificial intelligence (AI), particularly with the emerge of large language models (LLMs), I anticipate that spoken dialogue systems (SDS) will soon become the preferred and most widespread form of human-machine interaction, overtaking touch and type-based systems. Moreover, I foresee the next generation of dialogue systems shifting their focus towards an embodied setting, moving away from the traditional mobile-phone-based voice assistants. These dialogue-guided embodied agents are expected to have capabilities extending from performing simple household chores like cooking and cleaning to serving as assistants in shopping malls or as receptionists in banks

While being optimistic about the future, it's crucial for us as young researchers to have a thorough understanding of the major limitations of current spoken dialogue systems (SDS) and to focus on overcoming these barriers. In my view, the limited capabilities of current dialogue models to ground multimodal information, especially in the presence of lengthy and sophisticated linguistic contexts, represent a significant obstacle to the progress of SDS.

Furthermore, the data intensive nature of current visual-language models is a key factor that hinders the adaptations of SDS for multi-modal settings. However, the emergence of LLMs and Multimodal LLMs, which can be fine-tuned with limited amount of data, offers a promising avenue for overcoming these challenges.

3 Suggested Topics for Discussion

- How can multimodal dialogue systems benefit from large language models (LLMs)?
- What are the challenges of using LLMs as annotators or pseudo-label generators for unimodal and multimodal dialogues?
- How can knowledge distillation from LLMs contribute to building generalized multimodal dialogue systems?
- End-to-end or Modular: Should we reconsider multimodal dialogue architectures in the era of LLMs?
- How can we use function-calling abilities of LLMs to build multimodal conversational agents?
- How can we develop a multimodal Large Language Model capable of multi-turn dialogues across multiple modalities?

References

Yirong Chen, Ya Li, Tao Wang, Xiaofen Xing, Xiangmin Xu, Quan Liu, Cong Liu, and Guoping Hu. 2023. Exploring prompt-based multi-task learning for multimodal dialog state tracking and immersive multimodal conversation. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Proceedings of The Eleventh Dialog System Technology Challenge*. Association for Computational Linguistics, Prague, Czech Republic, pages 1–8. https://aclanthology.org/2023.dstc-1.1.

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. 2021. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*.

Bhathiya Hemanthage, Christian Dondrup, Phil Bartie, and Oliver Lemon. 2023. SimpleMTOD: A simple language model for multimodal task-oriented dialogue with symbolic scene representation. In Maxime Amblard and Ellen Breitholtz, editors, *Proceedings of the 15th International Conference on Computational Semantics*. Association for Computational Linguistics, Nancy, France, pages 293–304. https://aclanthology.org/2023.iwcs-1.31.

Bhathiya Hemanthage, Christian Dondrup, Hakan Bilen, and Oliver Lemon. 2024. Divide and conquer: Rethinking ambiguous candidate identification in multimodal dialogues with pseudo-labelling. In *25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 4903–4912. https://doi.org/10.18653/v1/2021.emnlp-main.401.

Yuxing Long, Huibin Zhang, Binyuan Hui, Zhenglu Yang, Caixia Yuan, Xiaojie Wang, Fei Huang, and Yongbin Li. 2023. Improving situated conversational agents with step-by-step multi-modal logic reasoning. In Yun-Nung Chen, Paul Crook, Michel Galley, Sarik Ghazarian, Chulaka Gunasekara, Raghav Gupta, Behnam Hedayatnia, Satwik Kottur, Seungwhan Moon, and Chen Zhang, editors, *Proceedings of The Eleventh Dialog System Technology Challenge*. Association for Computational Linguistics, Prague, Czech Republic, pages 15–24. https://aclanthology.org/2023.dstc-1.3.

Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. 2016. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, pages 792–807.

B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision (ICCV)*. pages 2641–2649. https://doi.org/10.1109/ICCV.2015.303.

Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, pages 69–85.

Biographical sketch

Bhathiya Hemanthage is a PhD student at the Edinburgh Centre for Robotics Centre for Doctoral Training, supervised by Prof. Oliver Lemon, Dr. Christian Dondrup, and Dr. Phil Bartie of Heriot-Watt University and Dr. Hakan Bilen from University of Edinburgh. His current research focuses on Generalized Visual-Language Grounding with Complex Language Context. He holds a Master's (by research) degree from University of Moratuwa, Sri Lanka. His Master's thesis supervised by Dr. Uthayasanker Thayasivam was on Dialogue State Tracking for Low Resourced Languages. Bhathiya has several years of experience as a Senior Software Engineer.