Developing a Mobile Learning and Tour-guiding System based on Generative AI Approach:

Fu-Kai Kuo[1] and Mu-Yen Chen[2][0000-0002-3945-4363]

[1] National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan (R.O.C)
N97111022@gs.ncku.edu.tw
[2] National Cheng Kung University, No.1, University Road, Tainan City 701, Taiwan (R.O.C)
mychen119@gs.ncku.edu.tw

Abstract. This paper presents a novel multimodal large language model architecture that operates flawlessly offline with low hardware requirements. It utilizes large language models (LLMs) as the brain, integrates the You Only Look Once (YOLO) object recognition model as the visual decoder, and incorporates auditory decoders and speech components to achieve capabilities of listening, speaking, reading, and thinking. Implemented on the compact Nvidia Jetson Orin, the system enables real-time interaction with real-world environments. Using the Tainan Confucius Temple as a case study, the system's effectiveness in mobile learning and digital tour guiding domains was explored and evaluated, demonstrating significant impacts on user experience and outcomes.

Keywords: Embedded Systems, Multimodal Large Language Model, Yolo, Edge Computing, Mobile Learning, Tour-guiding System.

1 Introduction

In recent years, large language models (LLMs) like ChatGPT[1] and LLaMA[2] have revolutionized our understanding and generation of natural language. With their profound linguistic comprehension, human-like text generation, contextual awareness, and powerful problem-solving capabilities, these models hold invaluable potential across various domains, such as search engines[3], customer support[4], and translation[5]. Among these, Multimodal Large Language Models (MLLMs)[6] have emerged as a novel focal point, leveraging the robust capabilities of LLMs to perform multimodal tasks like image analysis[7] and scene description[7], indicating potential pathways towards Artificial General Intelligence (AGI). However, there is currently a scarcity of research on multimodal large language models specifically designed for use with offline, low-resource edge AI devices. Therefore, this study introduces an innovative system framework to bridge this gap.

Concurrently, a variety of emerging information technologies have recently been applied to learning environments, particularly in mobile learning[8-10] and digital tour guiding[11]. These technologies offer visitors a diverse and convenient range of learning and tour-guiding services. Such systems are lightweight, portable, and personalized, allowing users to control the sequence and direction of their learning[10], coinciding with the devices used in this study. Studies using portable generative AI systems for this purpose are relatively scarce, so this study explores the educational effectiveness and usage intention of this innovative system in mobile learning and digital tour guiding domains, hoping that users can have a more profound travel experience through the system. Moreover, its offline usability characteristic allows the system to be deployed in any environment.

2 Materials and Methods

This paper introduces a new system framework that employs LLMs as the brain[6, 7], the object recognition model You Only Look Once (YOLO)[12] as the visual decoder, OpenAI's Whisper as the auditory decoder, and Linux's built-in Espeak as the system's voice, achieving capabilities of listening, speaking, reading, and thinking. Implemented on the compact Nvidia Jetson Orin, the system meets low hardware requirements and operates flawlessly offline, independent of network connectivity. Once pre-trained for specific contexts, the system offers endless possibilities for applications in new domains. By simply training the LLM and YOLO for the new domain, rapid deployment is facilitated. Utilizing Retrieval-Augmented Generation (RAG) technology[13], the LLM learns domain specific knowledge efficiently, without the need for traditional, time-consuming fine-tuning. Training YOLO with domain-specific images and annotations bestows the system with object recognition capabilities for the respective domain.

This paper examined the potential of an innovative system framework for mobile learning and digital tour guiding, focusing on Tainan Confucius Temple. Targeting four structures over 300 years old within the temple, the system was trained using the YOLO algorithm for object recognition and the RAG technique for domain-specific knowledge acquisition. On-site experiments invited visitors to engage with the system, which provided interactive information about the buildings upon detection. Visitors were divided into an experimental group, using the system for learning and tour guiding, and a control group, which participated in traditional guided tours.

3 Conclusion

The questionnaire in this study used quizzes to compare the learning outcomes of the experimental and control groups before and after the study. Additionally, scales were employed to investigate user intentions regarding the system, utilizing a modified AI Device Use Acceptance (AIDUA) model[14].

In essence, this study proposes an innovative multimodal large language model framework with low hardware requirements that interacts in real-time with real-world scenarios, evaluating user experiences and outcomes in mobile learning and digital tour guiding domains.

References

1. Ray, P.P., *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope.* Internet of Things and Cyber-Physical Systems, 2023.
2. Touvron, H., et al., *Llama 2: Open foundation and fine-tuned chat models.* arXiv preprint arXiv:2307.09288, 2023.
3. Caramancion, K.M., *Large Language Models vs. Search Engines: Evaluating User Preferences Across Varied Information Retrieval Scenarios.* arXiv preprint arXiv:2401.05761, 2024.
4. Lakhani, A., *Enhancing Customer Service with ChatGPT Transforming the Way Businesses Interact with Customers.* 2023.
5. Lu, Q., et al., *Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt.* 2023.
6. Wang, Y., et al., *Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning.* arXiv preprint arXiv:2401.06805, 2024.
7. Zhu, D., et al., *Minigpt-4: Enhancing vision-language understanding with advanced large language models.* arXiv preprint arXiv:2304.10592, 2023.
8. Chiang, T.H., S.J. Yang, and G.-J. Hwang, *An augmented reality-based mobile learning system to improve students' learning achievements and motivations in natural science inquiry activities.* Journal of Educational Technology & Society, 2014. **17**(4): p. 352-365.
9. Crompton, H. and D. Burke, *The use of mobile learning in higher education: A systematic review.* Computers & education, 2018. **123**: p. 53-64.
10. Chao, C.-M., *Factors determining the behavioral intention to use mobile learning: An application and extension of the UTAUT model.* Frontiers in psychology, 2019. **10**: p. 446627.
11. Aboelmagd, A., *Emerging Technology Trends in Tour Guiding: Virtual and Distance Tour Guiding.* مجلة كلية السياحة والفنادق. جامعة المنصورة, 2023. **13**(13): p. 341-370.
12. Jiang, P., et al., *A Review of Yolo algorithm developments.* Procedia computer science, 2022. **199**: p. 1066-1073.
13. Lewis, P., et al., *Retrieval-augmented generation for knowledge-intensive nlp tasks.* Advances in Neural Information Processing Systems, 2020. **33**: p. 9459-9474.
14. Gursoy, D., et al., *Consumers acceptance of artificially intelligent (AI) device use in service delivery.* International Journal of Information Management, 2019. **49**: p. 157-169.