Egocentric Perception of Walking Environments using an Interactive Vision-Language System

Haining Tan[1-3], Alex Mihailidis[1-3] , and Brokoslaw Laschowski[2-4]

[1]Computer Science, University of Toronto; [2]Robotics Institute, University of Toronto; [3]KITE Research Institute, Toronto Rehabilitation Institute; [4]Mechanical and Industrial Engineering, University of Toronto

Introduction

Large language models are able to provide detailed contextual understanding of a scene beyond what computer vision alone can provide, which can have significant implications for robotics and embodied intelligence. In this study, we developed a novel multi-modal vision-language system for egocentric visual perception, with an initial focus on real-world walking environments. Our new system addresses key limitations of previous research, including dynamic scene understanding beyond static object recognition, personalized user feedback for navigation, and real-time closed-loop human-AI interaction for improved decision-making.

Methods

We built and tested a number of state-of-the-art transformer-based encoder-decoder models, including the ViT vision encoder with GPT2 text encoder, and a generative image-to-text transformer with the CLIP vision encoder. We trained the models using over 43,000 image-text pairs of real-world environments that we adapted and labelled from the MSCOCO captions dataset for few-shot image captioning.

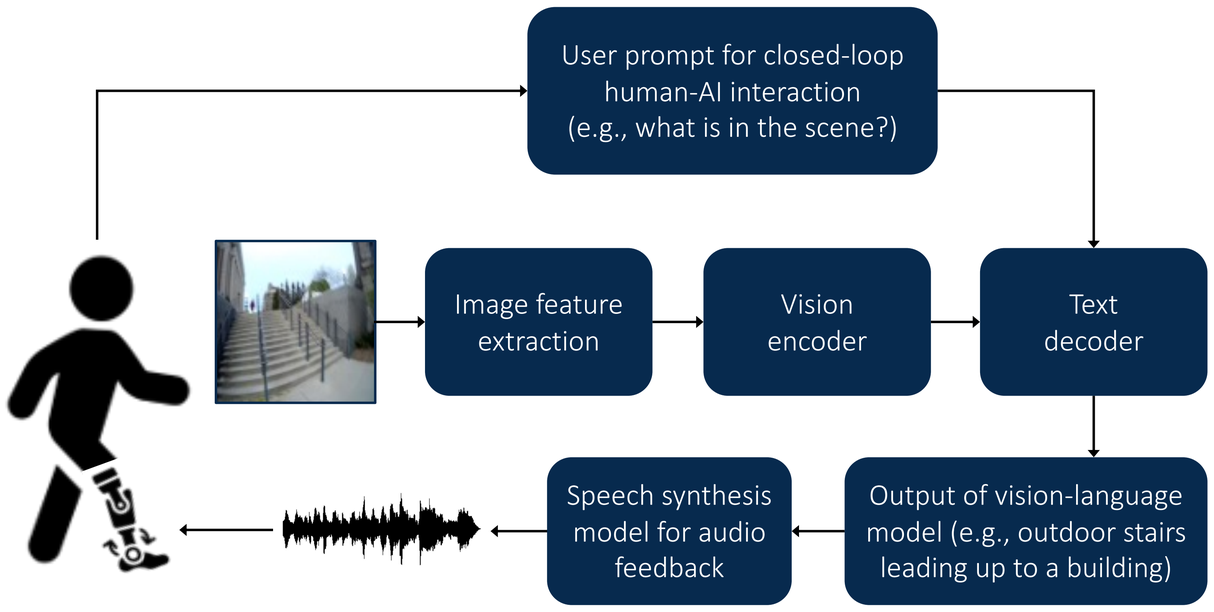


Figure 1: The end-to-end multi-modal pipeline for our vision-language system. Based on visual input of the walking environment and an optional user prompt, our system automatically generates text and audio outputs that describe the scene.

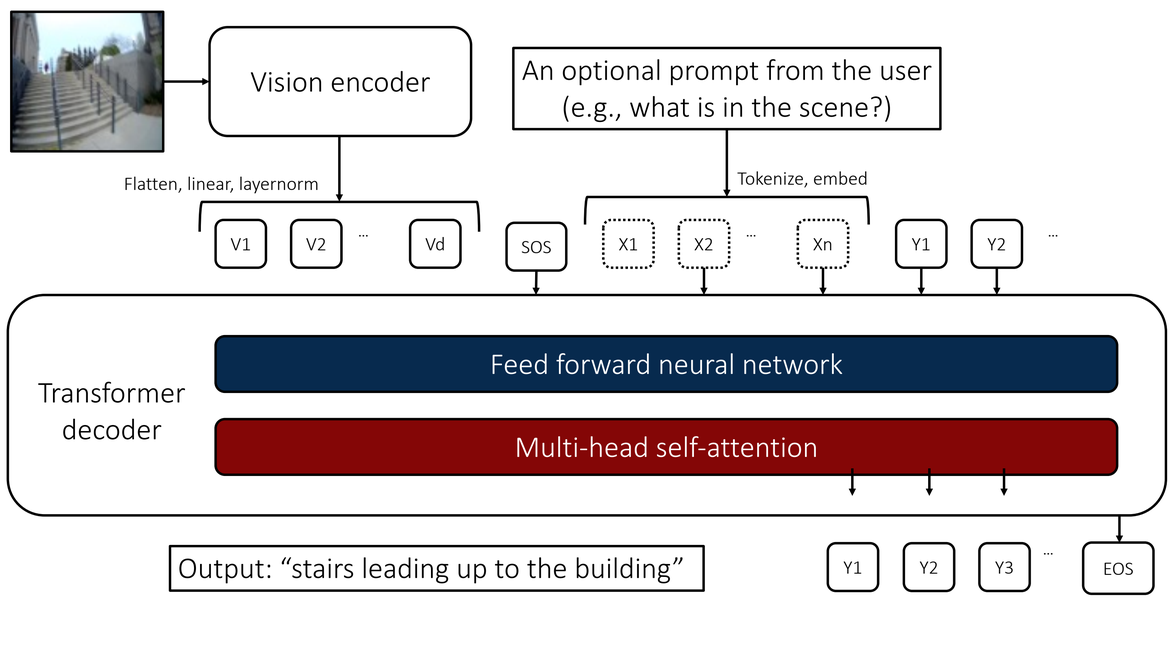


Figure 2: Our transformer-based encoder-decoder neural network architecture used for few-shot image captioning of real-world scenes, including an optional user prompt to tailor the generated image captions to the individual needs and preferences of the user.

We also custom-designed a speech synthesis module and a user interface to convert the generated image captions into speech for audio feedback to users. Our new system uniquely allows for user prompts to tailor the generated image captions to the individual needs and preferences of the user by integrating human cognition into the decision-making.

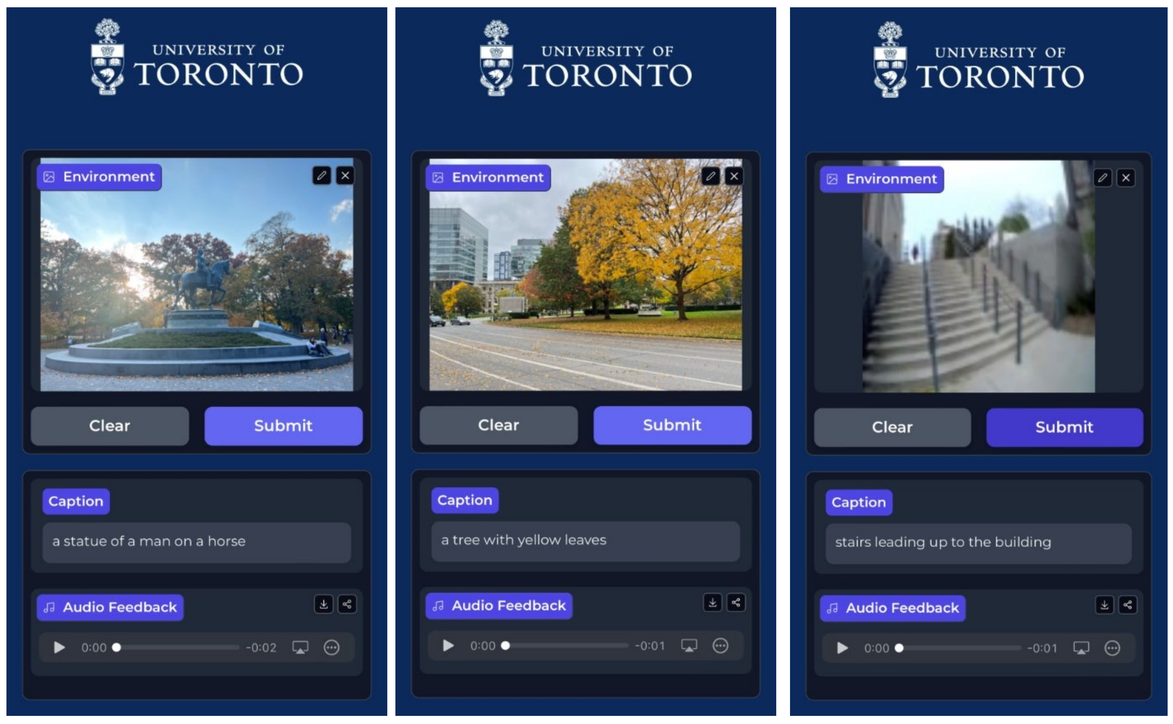


Figure 3: Our custom-designed interface to provide audio feedback to users describing the generated image captions by the vision-language models. Our interface also supports feedforward user prompts, thus allowing for closed-loop human-AI interactions and the ability to augment the decision-making of the vision-language models with human cognition.

Results

Our vision-language models were able to generate image captions with an average length of 10 words while achieving a high ROUGE-L score of 43.9% and low word error rate of 28.1% on unseen test images. Using cloud computing, our new system achieves an end-to-end processing time of 2.2 seconds to automatically generate image captions of real-world environments.

Our system variants	CPU	Nvidia T4 GPU	Nvidia A10G GPU
Text generation	7.5s	2.4s	2.2s
Text + audio generation	8.5s	3.0s	2.6s
Text generation with user prompts	30.2s	12.5s	11.3s
Text + audio generation with user prompts	31.1s	15.3s	12.4s

Table 1: The average processing speed of our four system variations, ranging from text generation only to both text and audio generation with additional user prompts.

Discussion

Our new multimodal large vision-language system can generate accurate and detailed descriptions of real-world scenes, which can be further augmented by user prompts. This innovative feature allows our generated image captions to be personalized to the individual needs and preferences of the user, thus optimizing the closed-loop interactions between the human and generative AI models for understanding and navigating of real-world walking environments.