Project Proposal: Scene Descriptions for the Visually Impaired CSCI 5541 NLP

Abbas Booshehrian, Alex Besch, Mohit Yadav, Ruolei Zeng
boosh002@umn.edu, besch040@umn.edu, yadav171@umn.edu, zeng0208@umn.edu
Sentimentals

1 Motivation and Problem definition

People with visual impairments face significant challenges in navigating their daily lives due to their inability to perceive and interpret their surroundings easily. The motivation behind this project is to empower these individuals by providing them with a sense of their environment using the latest advancements in deep learning, specifically Vision and Language Models. Recent developments in this field have demonstrated remarkable capabilities in understanding and generating human language.

Vision Language Models, such as OpenAI's ChatGPT-4o[1] and Google's Gemini[2] Multimodal, along with Large Language Models (LLMs) such as ChatGPT-3.5[3] and Claude Sonet[4], have made tremendous strides in recent years. These models can analyze images and generate natural language descriptions, offering new possibilities for assistive technology.

Our goal is to develop a system that can provide descriptive, language-based insights into an environment, with a focus on what is functionally relevant to the user. For instance, if a person is nearby, the system would prioritize describing that person rather than mentioning a distant building, as the former is more immediately useful for the user's understanding and interaction with their surroundings.

Therefore, the problem definition for our project is to develop a language descriptor system that processes streams of RGB-D camera images to extract and convey contextually relevant information, thereby aiding visually impaired individuals in understanding and navigating their environment. This system will harness the power of LLMs to create real-time, functional, and accessible descriptions, bridging the gap between visual perception and verbal communication for those who need it most.

2 Literature Review

Building on the success of VLMs and the capabilities of LLMs there is a push to create models based on 3D spatial understanding of scenes which have huge application in the field of robotics and autonomous vehicles. One of the leading work in the field in 3D-LLM: Injecting the 3D world into Large Language Models (Hong et al., 2023) in which researchers have created a model to take in 3D point cloud as input and perform a diverse set of 3D related tasks such as captioning, dense captioning, 3D question answering, task decomposition, 3D grounding, 3D-assisted dialog, navigation, and so on. The framework for this model is shown in Figure 1.

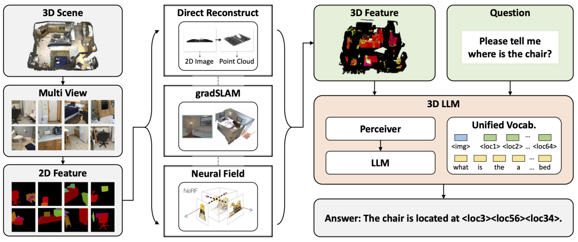


Figure 1: 3D-LLM Framework
(Hong et al., 2023)

Another important related work is the Contrastive Language-Image Pre-Training(CLIP) (Radford et al., 2021) network architecture that is trained on a variety of image-text pairs, although this work does not take the explicit 3D scene into consideration but this framework is widely used in Visual Language tasks. In particular, CLIP has demonstrated remarkable performance in tasks such as zero-shot classification, image retrieval, and multi-modal reasoning, making it a founda-

[1] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence
[2] https://www.gemini.google.com
[3] https://openai.com/chatgpt/overview
[4] https://www.anthropic.com/claude

tional model for integrating visual and linguistic information. The approach of CLIP to pair the images and text is shown in Figure 2 (Radford et al., 2021). The aforementioned approaches focus on 3D scene and Language binding(3D-LLM) or the task of visual image and language binding(CLIP). Our problem needs this type of understanding specifically designed to describe surrounding for a visually impaired.

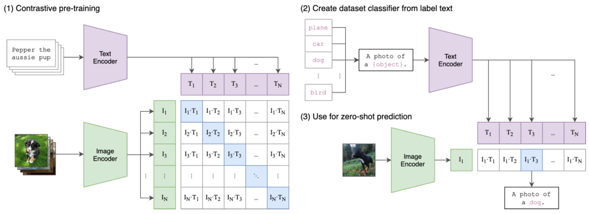

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T_1 T_2 T_3 ... T_N
I_1 I_2 I_3 ... I_N
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
$I_2 \cdot T_1$ $I_2 \cdot T_2$ $I_2 \cdot T_3$ ... $I_2 \cdot T_N$
$I_3 \cdot T_1$ $I_3 \cdot T_2$ $I_3 \cdot T_3$ ... $I_3 \cdot T_N$
$I_N \cdot T_1$ $I_N \cdot T_2$ $I_N \cdot T_3$ ... $I_N \cdot T_N$
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I_1
$I_1 \cdot T_1$ $I_1 \cdot T_2$ $I_1 \cdot T_3$ ... $I_1 \cdot T_N$
A photo of a dog.

Figure 2: CLIP Architecture

Although semantic rich scene understanding can help in many reasoning based methods, there is also a need for deep learning based methods for object detection and segmentation. These models don't provide a semantic understanding to the scene but can very accurately localize an object. The domain of object detection and labeling has improved immensely in recent years; You Only Look Once (YOLO) has been one of the main contributors in the real-time capability of object detection (Redmon et al., 2016). Fast YOLO, introduced in the same paper, achieved an impressive inference speed of 155 FPS. This capability is particularly crucial for real-time applications, such as assistive technologies, where immediate and accurate feedback is essential for ensuring user safety and effectiveness. A sample output from YOLO model is shown in Figure 3 (Redmon et al., 2016).

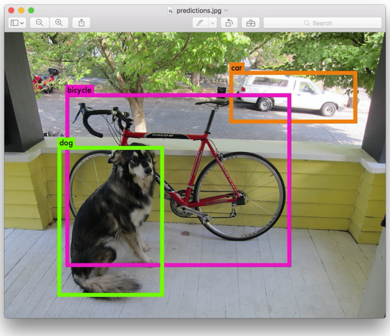


Figure 3: Sample Output from YOLO

Meta's Segment Anything Model[5] leads the image segmentation domain by providing segmentation masks for objects in an image. Without doubt, its ability to generalize across different object categories and provide fine-grained segmentation results has pushed the boundaries of what can be achieved in image segmentation tasks.

To provide a seamless user experience, one not only needs to understand the scene but should be able to convey the information to the human user in a meaningful and coherent manner. Large Language Models (LLMs) have been phenomenal in understanding and generating natural human language. Although there are many LLMs available, they all trace back to the seminal paper Attention is all you need (Vaswani et al., 2023). Vaswani et al. introduced a Transformer architecture (Figure 4) which may be considered a building block to the AI revolution we have seen in recent times [6].

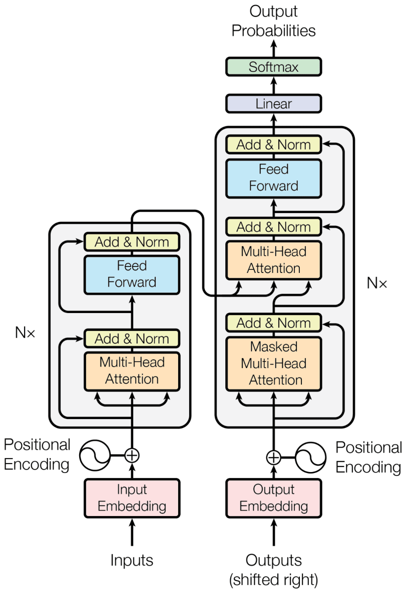


Figure 4: Transformer model architecture

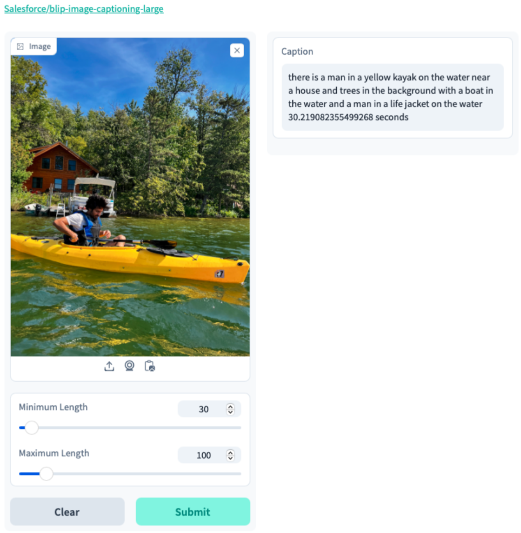
3 Execution Plan

We have observed efforts to integrate 3D scene information into LLMs with 3D LLMs and to combine vision with language in works such as CLIP. However, a limitation of these methods for our specific task is the inference time required. We tested the open source BLIP Image captioning Large model (Li et al., 2022) on Hugging Faces Space; the model took 30.21 secs for inference on

[5] https://segment-anything.com/

[6] https://machinelearningmastery.com/the-transformer-model/

the provided image snipped of input (output can be seen in Figure 5).



Caption

there is a man in a yellow kayak on the water near a house and trees in the background with a boat in the water and a man in a life jacket on the water 30.219082355499268 seconds

Figure 5: Sample BLIP inference

To address the issue of latency, we propose a novel approach for this task. Our method utilizes a 3D object detection model, which is significantly faster than Vision Language Models (VLMs) for detecting and labeling objects in an image. Our step-by-step approach is listed below:

1. We start by capturing an RGBD stream of the surroundings using an Intel RealSense L515 LiDAR-based camera.
2. We process the RGB images through an object detector to give bounding boxes and labels of the objects visible in the image.
3. The bounding boxes and labels are then attached to their depth values from the depth information obtained by the camera. Using the camera intrinsic parameters, we map each object to its 3D position with reference to the camera. This gives us a list of object labels with corresponding 3D locations in the scene.
4. This information, along with a timestamp, is fed into a large language model (LLM) to generate a coherent and relevant text output, which can then be read back to the user.

The proposed pipeline of our work is shown in Figure 6.

We argue that our specific application requires only this information to generate an output, which can assist visually impaired individuals in many scenarios without relying on time-consuming VLM models. By utilizing only an object detector and an LLM, we can generate captions that are equal to, if not better than, those produced by VLMs for our application.

3.1 Role Assignments

- Abbas: Aggregating image, depth, and time data to pass to LLM
- Alex: Image processing (Object detector, and distance attachment)
- Mohit: Image gathering, aggregating image, depth, and time data to pass to LLM
- Ruolei: Image processing (Object detector, and distance attachment)

4 Addressing Feedback

The major feedback received was to include a comparison of our model with an LLM in terms of inference time and the quality of output for relevance to our application, as well as the model's semantic understanding of the scene. We will incorporate this important feedback from the instructor team and present an analysis at the end of the project.

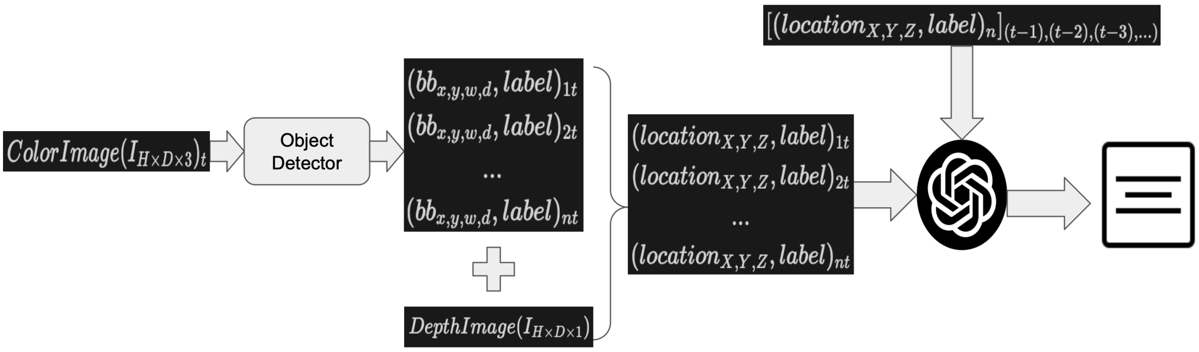

$ColorImage(I_{H\times D\times 3})_t$
Object Detector
$(bb_{x,y,w,d}, label)_{1t}$
$(bb_{x,y,w,d}, label)_{2t}$
...
$(bb_{x,y,w,d}, label)_{nt}$
$DepthImage(I_{H\times D\times 1})$
$(location_{X,Y,Z}, label)_{1t}$
$(location_{X,Y,Z}, label)_{2t}$
...
$(location_{X,Y,Z}, label)_{nt}$
$[(location_{X,Y,Z}, label)_n]_{(t-1),(t-2),(t-3),...)}$

Figure 6: Proposed Project Pipeline

References

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023. 3d-llm: Injecting the 3d world into large language models.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pretraining for unified vision-language understanding and generation.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. 2016. You only look once: Unified, real-time object detection.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.