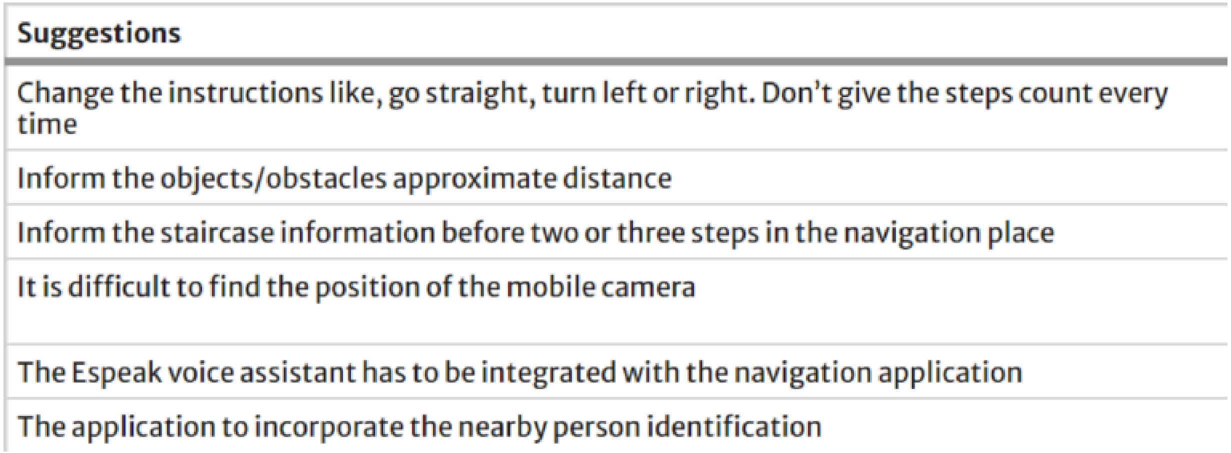
Deep Learning CSL4020
Indian Institute of Technology, Jodhpur

Problem statement

Design a deep learning model for Indoor Navigation for the Visually Impaired.

Solution Strategy

Initially, we had two distinct approaches to this problem: the former entailed providing the user with explicit guidance regarding the optimal route and requisite steps through a three-dimensional representation of the available scene, while the latter involved providing the user with information about the locations of obstacles, thereby allowing them to determine their preferred pathway autonomously. Upon researching more on the problem, we found a study that surveyed 50 visually impaired participants for this problem, and these were the suggestions given by the participants:



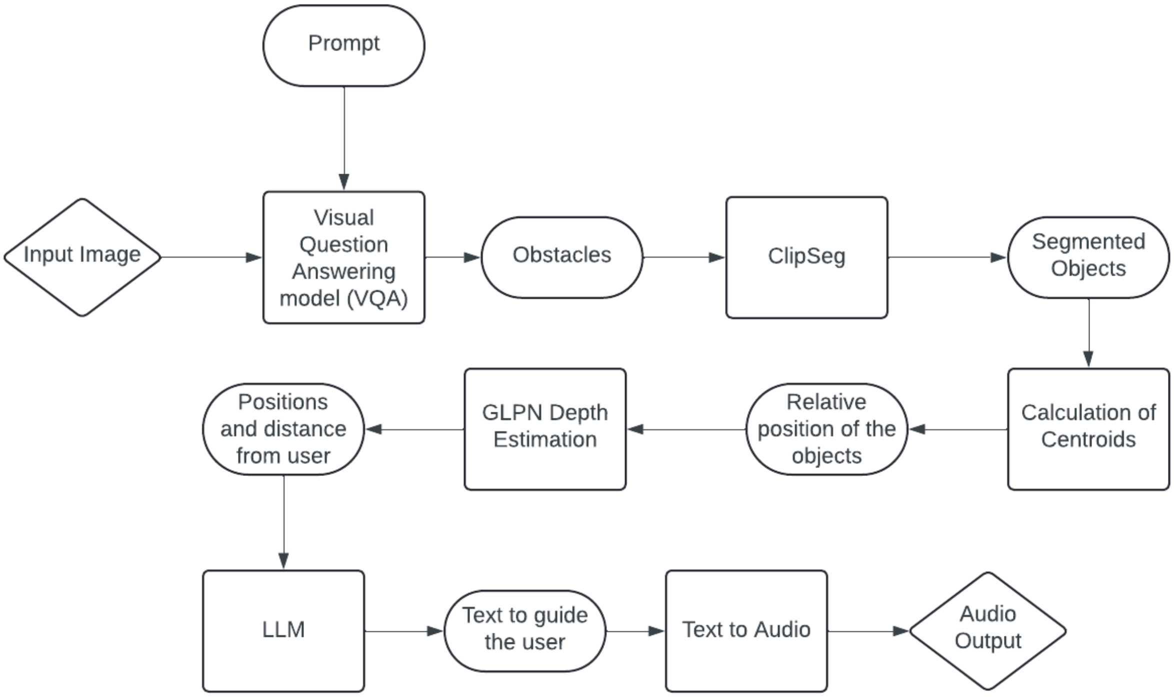
Suggestions
Change the instructions like, go straight, turn left or right. Don't give the steps count every time
Inform the objects/obstacles approximate distance
Inform the staircase information before two or three steps in the navigation place
It is difficult to find the position of the mobile camera
The Espeak voice assistant has to be integrated with the navigation application
The application to incorporate the nearby person identification

So, based on this, we decided to go with the second approach.

Initially, we took the image and used a ***Visual Question-answering model*** (VQA), namely ViLT, to generate a compilation of things detected in the image that could become potential obstacles. We obtained a roster of five entities using this model. Next, we used a segmentation model called ClipSeg to segment the objects acquired from the VQA model into distinct segments. Then, we computed the centroids of these segmented objects to determine their positions relative to the image center. This allowed us to inform the user about the obstacle's location in relation to their position - whether it was to their left, right, above, or below.

But there was still one problem: the presence of several obstacles. This raised the question of how the user could initially determine which obstacle to prioritize. To address this issue, we employed depth estimation models (GLPN) and a depth map to indicate to the user the object nearest to them and the farthest away.

After gathering all the necessary information, we utilized an LLM (Large Language Model) to generate the instructions required for assisting visually impaired individuals. Subsequently, we employed a text-to-speech library to deliver these instructions based on the insights from the earlier survey.



Dataset

We were originally asked to use the InLoc dataset, but it was very difficult to find smaller versions as it was depreciated. The only available version was very large (approx 400 GB), and we did not have the computational power to work with such a big dataset. So, we decided to explore alternative, more feasible options given our computational constraints. After careful consideration, we have selected the DAQUAR (Dialog-based Question Answering) dataset as the primary dataset for our project.

The DAQUAR dataset is a well-established resource for research in the field of visual question answering (VQA). It consists of a collection of images accompanied by natural language questions and their corresponding answers. The dataset was initially introduced in 2015 and has since been widely used in the research community. Compared to the larger MS COCO dataset, DAQUAR presents a more manageable size, making it more accessible for our current project.

Additionally, since we were already utilizing pre-trained models that had been trained on the MS COCO dataset, we decided to leverage the DAQUAR dataset to build upon our existing knowledge and expertise.

Some images from the DAQUAR dataset →

Major Innovations

In our project, we identified a critical gap in existing solutions: they predominantly offered users predefined paths, restricting their autonomy and freedom of exploration. Moreover, these solutions lacked dynamism, constraining users to static interactions devoid of real-time adaptability. *We decided to use a VQA model rather than the traditionally used Image captioning* *models, allowing the user to interact with the model and giving the users the ability to know what is present around them and giving them a dynamic interface to communicate with the model to get each and every insight of the scene.* We also incorporated ClipSeg and depth estimation models to provide the user with more information about the scene and let the user know the relative positions of each obstacle that could be in his path, allowing them to navigate around more accurately.

Location of Object:

- To provide a qualitative idea of the location of an object in an image relative to the viewer's eyes (camera), we can use the CLIPSeg model. This model takes the image as input and a label specifying the object we want to locate within the image. It then outputs a 2D array where the pixels corresponding to the object are assigned high values (appearing white), while the rest of the image is assigned low values (appearing black).
- To determine the direction of the object relative to the viewer's eyes, we can calculate the object's centroid. This is done by computing the mean of the indexes of the high-value pixels in the 2D array outputted by the CLIPSeg model. The centroid represents the approximate center of the object within the image.
- Next, to gauge the direction in which the viewer should move to approach the object, we calculate the center of the entire image, representing the current viewing perspective of the viewer. By determining the quadrant in which the object's centroid lies relative to the center of the image, we can infer the direction in which the viewer should move.
- For example, if the object's centroid lies in the top-left quadrant of the image center, it suggests that the object is located in the upper-left direction relative to the viewer's current position. Conversely, if the centroid is in the bottom-right quadrant, it indicates that the object is situated in the lower-right direction from the viewer's perspective.

Closeness of Object:

- We utilize a depth map generated using a GLPN (Global-Local Path Networks) model to estimate the object's distance within the image. This depth map represents the relative distances of objects within the scene captured by the camera. In the depth map, darker pixels correspond to objects closer to the camera, while lighter pixels represent objects farther away.
- Once we have the depth map, we can use the object's centroid obtained earlier to estimate its distance from the camera. We can extract the

corresponding depth value assigned to that particular pixel by locating the centroid within the depth map. This depth value indicates the relative distance of the object from the camera.

- However, to provide a relative estimation of the distance, we need to consider the distribution of depth values within the depth map. Typically, higher depth values (appearing white) correspond to objects that are at the greatest distance from the camera, while lower depth values (darker pixels) represent closer objects.
- Based on this distribution, we can categorize the depth values into different ranges to represent varying distances from the camera. For example, depth values falling within a certain range might indicate objects that are relatively close to the camera, while values in another range might represent objects at intermediate distances. Similarly, the highest depth values correspond to objects at the farthest distance.
- By categorizing the depth values in this manner, we can provide a qualitative estimation of the object's relative distance within the scene captured by the camera. This approach allows us to gauge the spatial relationship between the viewer and the object, enhancing our understanding of the scene's depth and perspective.

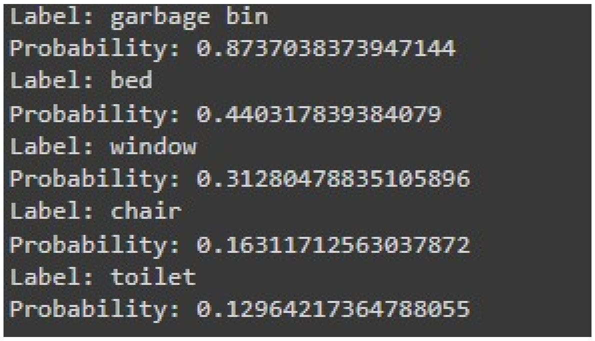
Contributions

Vidit Agrawal - VQA model and Image Captioning Literature Review
Vidit Garg - Depth estimation
Shreejan Kumar - Segmentation
Kushal Agarwal - LLM and audio processing

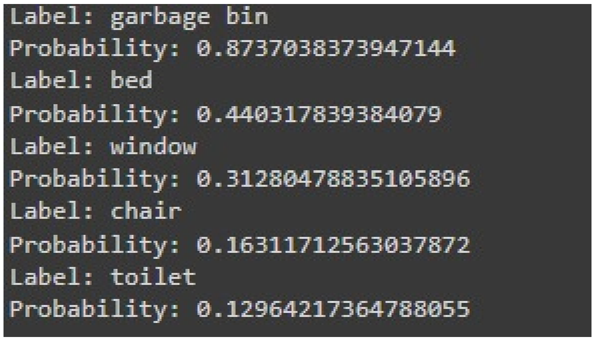
Results

Input Image:

Output of fine-tuned VQA model and VQA model designed from scratch:

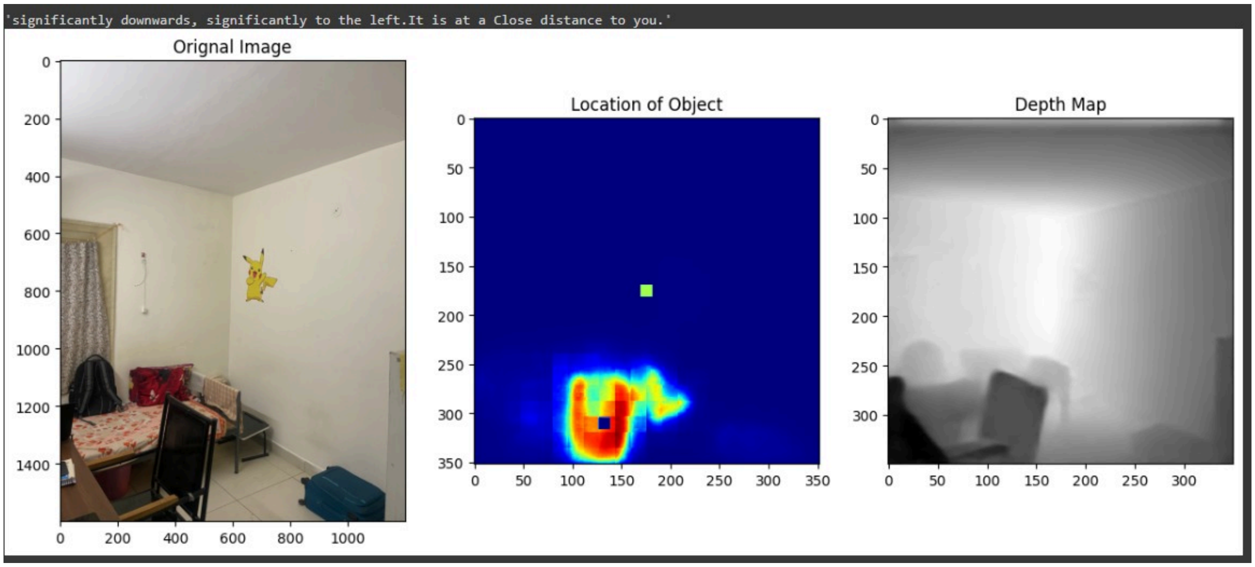


```
Label: garbage bin
Probability: 0.8737038373947144
Label: bed
Probability: 0.440317839384079
Label: window
Probability: 0.31280478835105896
Label: chair
Probability: 0.16311712563037872
Label: toilet
Probability: 0.12964217364788055
```



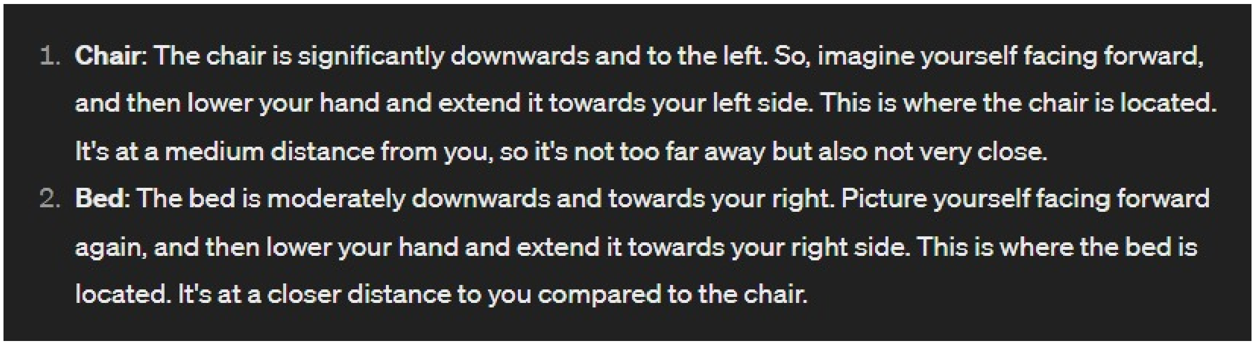
```
Label: garbage bin
Probability: 0.8737038373947144
Label: bed
Probability: 0.440317839384079
Label: window
Probability: 0.31280478835105896
Label: chair
Probability: 0.16311712563037872
Label: toilet
Probability: 0.12964217364788055
```

Segmentation of chair along with depth map and position of object:



```
'significantly downwards, significantly to the left.It is at a Close distance to you.'
```

Output from LLM:



1. **Chair**: The chair is significantly downwards and to the left. So, imagine yourself facing forward, and then lower your hand and extend it towards your left side. This is where the chair is located. It's at a medium distance from you, so it's not too far away but also not very close.
2. **Bed**: The bed is moderately downwards and towards your right. Picture yourself facing forward again, and then lower your hand and extend it towards your right side. This is where the bed is located. It's at a closer distance to you compared to the chair.

After this we will be using a Text to speech model and we will get an audio output that will help the visually impaired individual to get to know his surroundings.

Analysis of the Solution

The problem with the existing solution is the lack of quantitative estimation of the object's distance from the viewer and the absence of information about the object's dimensions. These limitations indeed hinder the accuracy of determining how close an object truly is, especially when considering the size and dimensions of the object.

Consideration of Object Size:
Accounting for the object's size is essential, as larger objects may have centroids that appear farther away, even if their closest boundaries are relatively close to the viewer. By considering the object's size, we can adjust our depth estimation approach to prioritize the proximity of the object's closest boundaries to the viewer rather than solely relying on the centroid position. To tackle this problem, we can create a 3D-dimensional space of the room and make the objects' bounding boxes. To create a 3D dimensional space, gather a dataset of images captured within the room where spatial understanding is required. These images should cover various perspectives and viewpoints within the room, showcasing different objects and their spatial arrangements. Training a deep learning model to create a 3D-dimensional representation of the room using the collected images. This involves learning the spatial layout, dimensions, and features of the room, including the positions of walls, furniture, and other objects. By combining the 3D-dimensional representation of the room with object bounding boxes and dimension estimates, the model gains a comprehensive understanding of the spatial layout and relationships between objects within the room.

Deploy the trained model in real time to perform spatial understanding tasks within the room. As the camera captures new images within the room, the model can generate 3D-dimensional representations, detect objects, estimate dimensions, and provide spatial insights on the fly.

Conclusion

Our project culminates in successfully creating a pipeline specifically tailored to meet the demands of visually impaired individuals. Our solution combines advanced technologies such as Visual Question-answering models, segmentation techniques, depth estimation algorithms, and natural language processing capabilities. This allows us to answer users' questions and provide a detailed description of the scene, including the presence and spatial orientation of objects and their distances.

This pipeline provides users with enhanced independence, as they may confidently and clearly traverse their environment, equipped with up-to-date information about any barriers. Our idea represents a substantial development in assistive technologies by combining artificial intelligence with human interaction. This innovation paves the path for a more inclusive and accessible environment for visually impaired individuals.

References

ViLT - https://arxiv.org/abs/2102.03334
ClipSeg - https://huggingface.co/docs/transformers/en/model_doc/clipseg
GLPN - https://huggingface.co/docs/transformers/main/en/model_doc/glpn

Team Members :
Vidit Agrawal (B21AI058)
Shreejan Kumar(B21EE088)
Vidit Garg(B21AI045)
Kushal Agarwal(B21BB017)