Perceiv.

Sounding Vision - AI powered Visual Assistant
Perceiv. is a multimodal visual assistant which is designed to help visually impaired individuals better understand their surroundings so they can more easily navigate and interact with unseen world. They would have an easier time understanding surrounding objects and places through vocal conversations with the visual assistant. The mix of image captioning, conversational AI, speech recognition, and object localization, transforms visual information to auditory descriptions resulting in a conversational flow.

Perceiv utilised four fundamental technologies which are:

BLIP: Bootstrapping Language-Image Pre-training
LLAMA2: Large Language Model Meta AI 2
CLIP: Contrastive Language-Image Pre-Training
GTTS & GSTT: Google Speech-to-Text and Google Text-to-Speech
Architectural Overview

Perceiv has a modular architecture as it contains several different components each performing specific tasks and each component works with the other within the pipeline to provide the user with the desired output.

The workflow begins with generating image captions and ends in either a conversational piece or a summary of the input image. The response is based on the type of query the user has input, either an image based query or a broader query that is unrelated to the image.

Key Modules:

Image Captioning: LORA fine tuned BLIP generates a text description of the input image.
Query Understanding: User queries (audio input) are processed using Whisper for speech-to-text conversion.
Dynamic Routing: Based on the type of query the input will either be fed to LLAMA2 or CLIP to obtain appropriate responses
Response Generation: The output received either from LLAMA2 or CLIP, is converted to audio via Whisper

Detailed Workflow

Image Input and Caption Generation:
User captures image using their preferred smart device.
Captured image is sent as an input to BLIP for processing, which generates a detailed and concise caption describing the objects, relationships and context within the scene.

(ii) User Interaction:
The user interacts with the assistant via speech.
User can either ask image specific questions or questions that are completely unrelated to the image.
Whisper converts the speech input into text.

(iii) Query Routing:
Image specific query
Queries that are related to the input image. For example, “Where is the chair in the image?”, such a query will be routed to CLIP and not LLAMA2.
General knowledge query
Queries that are completed unrelated to the input image. For example, “Who invented the sandwich?”, such a query will be routed to LLAMA2 and not CLIP.

(iv) Response generation and output:

The response either from CLIP or LLAMA2 will be sent to Whisper to convert it from text to speech.
The final audio response is then played out the user.

(v) User feedback loop:
The user can further interact with the assistant about the input image or other broader questions unrelated to the image.

Applications and Use cases

Accessibility for the Visually Impaired:
The assistant converts the visual information into auditory descriptions and answers question about the surrounding environment based on the user's query. Enables the users to "see" their surroundings via detailed captions and conversational pieces.

Educational Tools:
It can facilitate learning by providing detailed descriptions and interactive QnA about objects or scenes.

Industrial Assistance:
It helps in identifying and localising components within complex systems like machinery or vehicle engine bays.

Augmented Reality (AR):
The assistant can be integrated with AR systems to provide real time annotations and explanations about the surrounding environments and objects.