Edge-Based Vision Assistant

Aditya Shetty, William Borom

Advisor: Roger Iyengar
15821 - Fall 2021

Carnegie Mellon University
School of Computer Science

Overview

Edge-Based Vision Assistant acts a tool that visually impaired users can use to assess an unknown environment in front of them. It uses machine learning models run on a remote server to make inferences about a given scene and determine the locations of objects within the scene. Unlike competitor products, Edge-Based Vision Assistant uses completely open source tools and has toggleable functionality through physical motion (amenable to blind users).

Background and Related Work

Microsoft Seeing AI
- Scene
- Color
- Handwriting
- Text
- Barcodes
- Familiar People
- Currency

Google Lookout
- Grocery Labels
- Various Languages
- Document Scanning

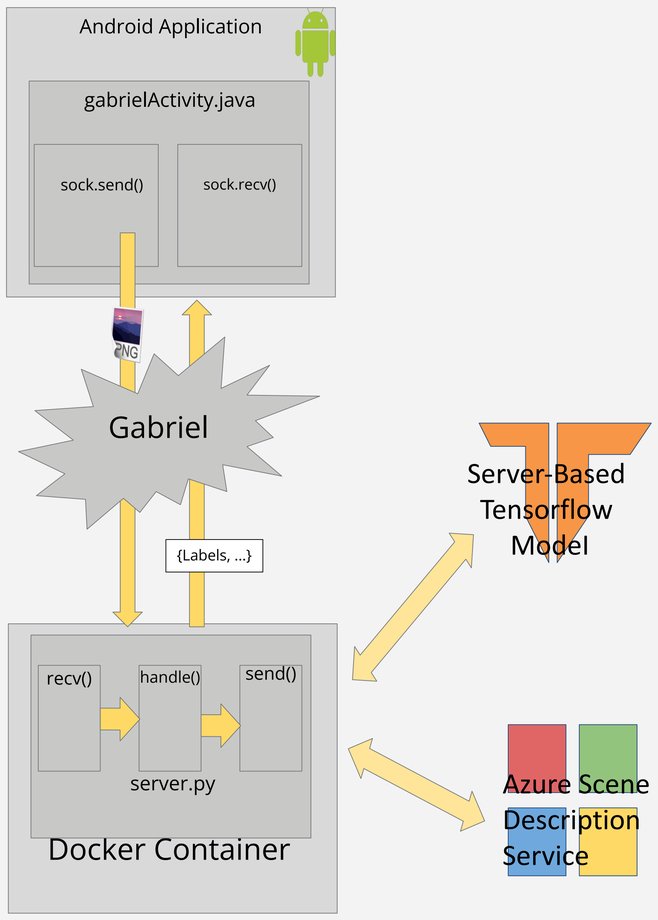
Im2Text Show and Tell
- Static Image Scene Description

System Architecture

Client Side
- Android Application creates Gabriel Activity
- Activity initializes TTS and other sensor readers including camera capture
- Opens Websocket connection with cloudlet

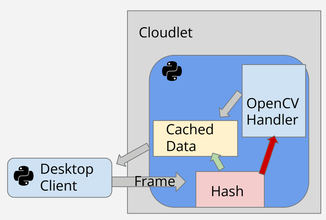
Gabriel
- Framework that enables traffic control from client to cloudlet
- via tokens

Server Side
- Running within Docker
- Receives Image Bytes + params
- Uses OpenCV for object detection
- For scene description can use either local TF model or Azure
- Service



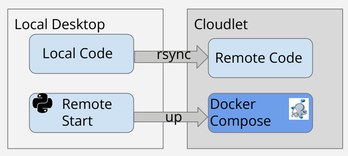
Features

Perceptual Hashing



- Reduces unnecessary computation on the server side
- If the perceptual hash of a frame is similar to a certain threshold of the last frame,the current frame is not processed
- If hash sufficiently different, trigger new scene description

Deployment

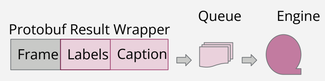


- Remote start script sends new server code to Gabriel Server
- Server Process runs in a Docker Container

Object Detection and Scene Inference

- Bounding box and object labels determined with YOLO
- Scene Inference is performed with modified Show-and-Tell model [arXiv:1411.4555 [cs.CV]]
- We trained this model on 80,000 images

Text To Speech



- Scene Inference is performed with Show-and-Tell model
- We trained this model on 80,000 images

Use of Accelerometer

- Eliminates the need for manual button press (not amenable to blind users)
- Toggles functionality

Proof of Concept

"You are looking at a stainless steel refrigerator freezer with in kitchen."

"From top to bottom I see refrigerator sink, laptop.

From left to right I see laptop, refrigerator, sink."