# Visual Question Answering

Feiyu Wang, Yifei Zhu, Fei Gao

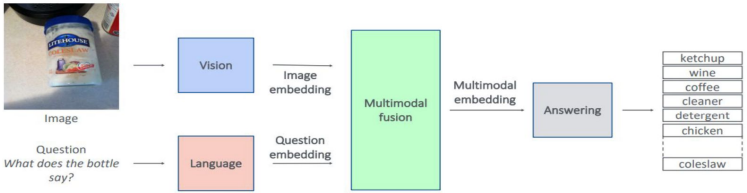
BOSTON UNIVERSITY

GitHub

## Introduction

The product of VQA is to set for helping the visually impaired identify daily necessities and give them essential instructions.

*** Identify objects and Answering questions**
***Text to Voice function**
***High Accuracy for required environment**



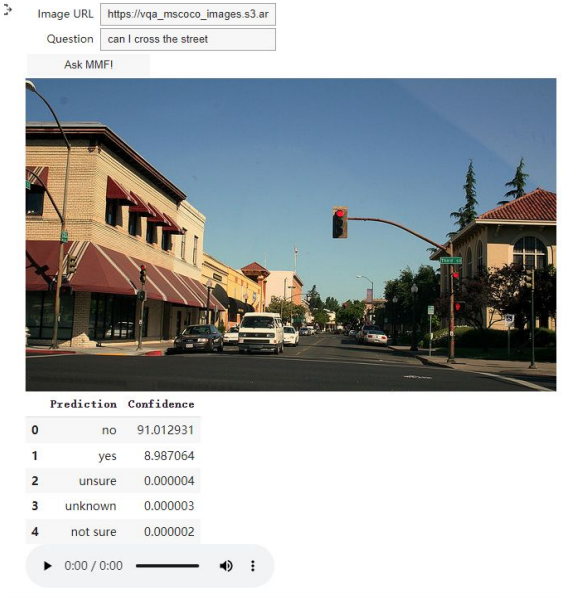

## Voice Function

**Text To Speech:** gTTS. - a screen reader application developed by Google for the Android operating system. It powers applications to read aloud (speak) the text on the screen with support for many languages.
**Speech To Text:** Silero Speech-To-Text Models from Silero AI Team. A set of compact enterprise-grade pre-trained STT Models for multiple languages. This models are robust to a variety of dialects, codecs, domains, noises, lower sampling rates.

## Demo

Through voice function, blind people are able to get assistance.

Red light scene.
**Blind people:"Can I across the street?"**
**VQA product: "No."**




## Model

**MoViE+MCAN model**: formerly known as Pythia. The MMF uses the same based technologies - Bottom up and top down to analysis imagines and response with answers by the question raised.

## Dataset

**VQA 2.0:** a new dataset containing open-ended questions about images. These questions require an understanding of vision, language and commonsense knowledge to answer.

| Model | Dataset | Metric | Notes |
|---|---|---|---|
| Pythia | vqa2 (train+val) | test-dev accuracy - 68.31% | Can be easily pushed to 69.2% |