Evision

A web app designed for the low-vision community

Aayusha Odari[1*], *Aditi Kharel*[2*], *Mamata Maharjan*[3*]

[1,2,3]**Department of Electronics and Computer Engineering, Institute of Engineering, Pulchowk Campus, Nepal**

[1,2,3] (077bct006.aayusha , 077bei008.aditi, 077bct043.mamata)@pcampus.edu.np

Evision is an application designed to empower the low-vision community of Nepal. It implements the LLaVa model for extracting descriptions and a Text-to-Speech (TTS) system based on Tacotron2 for language articulation. When users click the image from the application, Evision extracts the features of the scene and articulates them in Nepali, thereby enhancing accessibility.

1 Background

Before arriving at this concept, we analyzed the problems faced by the low-vision community of Nepal. Through an interview and observations, we got to know the difficulties faced by individuals with visual impairments in navigating their surroundings, accessing information, and engaging with the world around them. We came to find that voice is everything to them. Motivated by these findings, our objective became clear: to develop a solution that empowers individuals to perceive the world as it truly is. Thus, our application Evision was born.

2 Introduction

Visual impairment hinders the independence of the people, creating a barrier to engage with the world around them. Evision aims to help individuals to perceive their surroundings. Through the application, users can capture the surroundings from their smartphone's camera and receive a description of the scene in Nepali language. This enables them to understand the environment through auditory feedback. Thus, it is to be believed that Evision seeks to improve the quality of life for visually impaired individuals in Nepal.

3 Methodology

3.1 Image Description Extraction

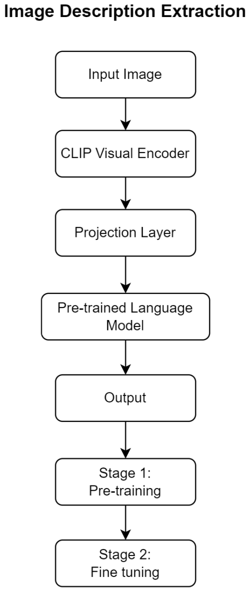

Figure 1: Block diagram for Image Description Extraction

For extracting the image description, we have used LLaVA: Large Language and Vision Assistant, an end-to-end trained large multimodal model that connects a vision encoder and an LLM. [1] LLaVA's architecture integrates a pre-trained language model, Vicuna, with a CLIP visual encoder to process both textual and visual inputs effectively. The input image undergoes feature extraction by a pre-trained CLIP visual encoder. These features are then projected using a trainable projection layer. This projection process generates language tokens that are fed into the pre-trained language model to generate the predicted output. The training of LLaVA involves fine-tuning the pre-trained language model and visual encoder on task-specific data to optimize performance for multimodal tasks

3.2 Text-to-Speech

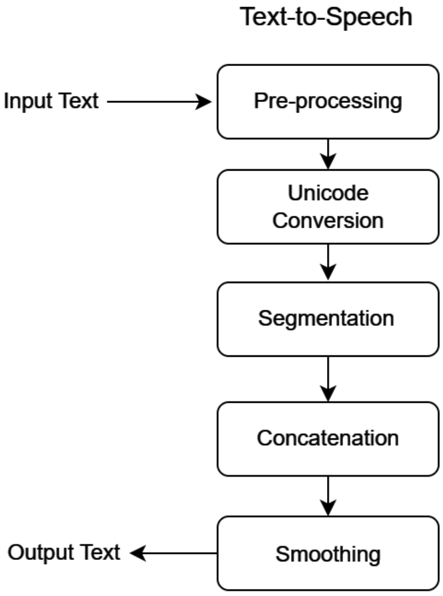

Figure 2: Text to Speech Formulation

In the development process, we utilized TensorFlow and Flask for handling the aspects of the Text-to-Speech functionality in Evision. The textual data is subjected to preprocessing to ready it for synthesis. The text is transformed into Unicode for character representation, with segmentation, concatenation, and smoothing applied as needed to ensure natural-sounding speech. Evision is supported by a Tacotron2-based Text-To-Speech Synthesis Model for Speech Synthesis.

Additionally, audio resources were obtained from the existing project "Nepali Text-to-Speech Synthesis using Tacotron2 for Melspectrogram Generation". [2]

4 Features

- Camera integration: It incorporates with smartphone's camera for capturing surrounding.
- Nepali Language: It is the first-ever application that describes the scene features in the Nepali language.
- Natural Voice: Evision ensures a natural voice in the output.
- Accessibility: The application is easy to use and accessible to all.

5 Conclusion

Evision represents a significant step forward in addressing the challenges faced by the low-vision community in Nepal. This application is especially useful for those who may have limited proficiency in English or prefer to interact in their native language. Through localization, it helps users feel connected to their culture and community. With its user-friendly interface and innovative features, Evision is sure to improve the quality of life for the visually impaired population in Nepal.

References

[1] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," 2023.

[2] S. Khadka, R. G.C., P. Paudel, and R. Shah. Nepali text-to-speech synthesis using tacotron2 for melspectrogram generation. Available: https://gitlab.com/shrutiaudio/shrutiaudio.