A VOICE BASED IMAGE CAPTION GENERATOR USING DEEP LEARNING

Project Reference No.: *47S_BE_2114*

College ***:*** *Jain College of Engineering and Technology, Hubballi*
Branch ***:*** *Department of Computer Science and Engineering*
Guide(s) ***:*** *Prof. Trupti Thite and Mr. Venkatesh M E*
Student(S) ***:*** *Ms. Pooja*
Ms. Renuka
Ms. Tejasvini Tejappanavar
Ms. Vajreshwari

Keywords:

Image caption generator, Flickr 8K dataset, LSTM, gTTS engine

Introduction:

In recent years Deep learning is one of the most popular approaches in Machine Learning and artificial intelligence.

It is a machine learning Technique inspired by the Human brain, it uses the algorithm like convolutional neural network, recurrent neural network, long short-term memory etc., where there are many developments had already made for partially visually impaired people.

A voice-based Image caption generator is used to identify the objects and information present in the image, which could improve the lives of partially visually impaired people.

Using CNN and LSTM together can be best fit for this project because LSTM is similar to RNN, and the RNN algorithm is depending on the LSTM because it's having the feedback connectivity and also LSTM process the entire sequence of data.

Scope / Objectives of the project:

In order to develop the system following objectives are set-

The main objective of the project is to develop a system capable of generating descriptive caption of images.

To develop the system that facilitates the dual communication channel by combining an image description and voice-based generator.

To design and develop a system for partially visionless users and for untutored people.

To integrate both visual and auditory components in order to enhance the user experience.

Methodology:

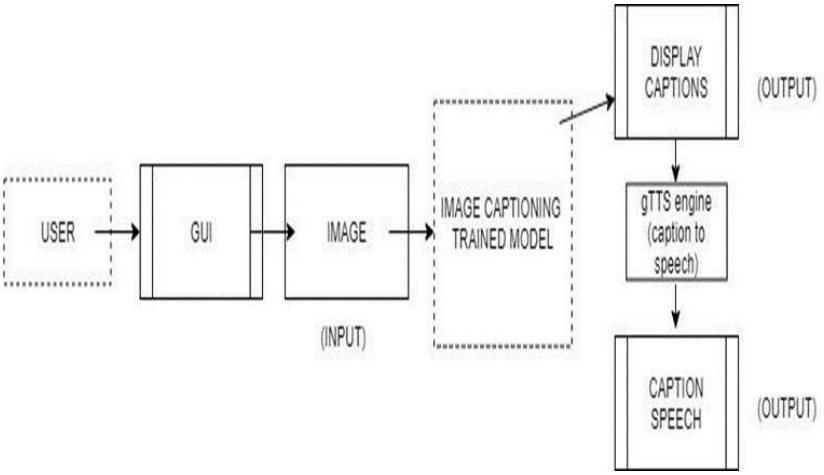


Fig1.Block diagram.

To implement a python web application using GUI with Machine/Deep learning techniques.

- The model generates image captions using merging model along with RNN and LSTM.
- This model is trained on Flickr-8k dataset which generates the most preferable captions and also has a mechanism to converts the generated caption into speech using gTTS engine ultimately helping out the visually impaired.

DATA SET

There are multiple data set which are open and are available for training models –

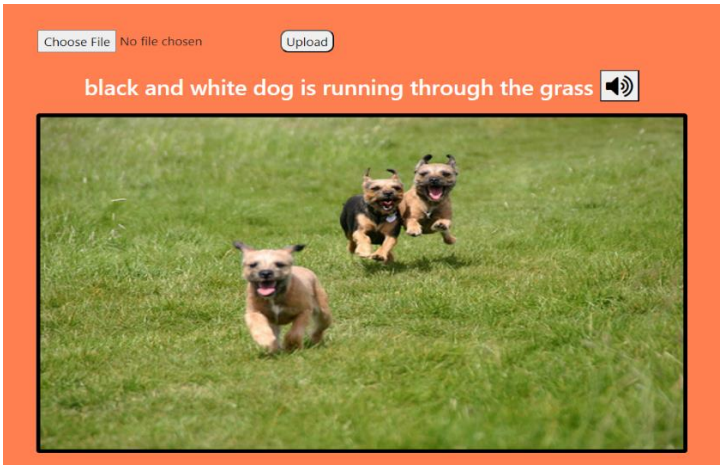
- Flickr 8k (which contains 8000 images)
- MS COCO (which contains 180000 images)
- Flicker 30K (which contains 30000 images) and etc.

For our preliminary use we have taken the Flickr 8k dataset. This dataset contains 8000 images each having 5 captions which are divided as 6000 images for training sets, 1000 for development sets and another 1000 for testing set

Results and Conclusion:

- The system also generates the speech output from the captions obtained.
- Voice based image caption generator has been developed using a CNN-LSTM model.
- Main key aspects of our project to note, the proposed model not only depends on the dataset, the proposed model is trained for testing the user input, so that it can predict the descriptions from the external images.

- Out dataset consists of 8091 images. The proposed model is required to be trained on huge dataset that contains more than 10,000 images for achieving a better accuracy.
- The functional requirements of the system ensure that it can perform its key features.
- While the non-functional requirements ensure that the system is reliable, secure, and user-friendly



Innovations in the project:

- Development of a system that not only accurately transcribes spoken descriptions of images but also generates descriptive captions that are contextually relevant and linguistically fluent.

- The system involves incorporating real-time object detection and scene understanding capabilities into the system, allowing users to interactively describe specific elements within an image and receive instant, detailed captions for those elements.

- This project aims to enhance accessibility and user experience by enabling seamless interaction with images through natural language commands and descriptions.

Scope for future work:

- Future work will focus on enhancing the integration of information from different modalities, such as text, image, and audio, to improve the accuracy and contextuality of generated captions.

- Advancements are expected in developing models capable of capturing intricate details, relationships, and spatial arrangements within images, enabling more nuanced and comprehensive descriptions.

- Efforts will be made to create interfaces that adapt to individual user preferences and provide interactive feedback mechanisms, fostering greater user engagement and satisfaction.

- Addressing ethical concerns such as bias mitigation, fairness, and inclusivity will remain crucial to ensure that captioning systems are developed and deployed responsibly, respecting diverse perspectives and user needs.