International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 04 | April - 2023 Impact Factor: 8.176 ISSN: 2582-3930

CAPTIONBOT FOR ASSSISTIVE VISION

[1]Ms. E. Padma, [2]Y. Jayanth Kumar, [3]U. Pradeep Chowdary

[1]Assistant Professor, [2]Graduate Student, [3]Graduate student

Dept. of Computer Science Engineering,

SCSVMV (Deemed to be University), Enathur, India

Abstract: In view of the recent flow of work in intelligent robotics, to our knowledge, the results from this research field have scarcely been applied to mitigate sensorial, motor and cognitive damage for humans. We believe that such research, in particular technique of learning from demonstration in imitation learning, is well suited to addressing the problem during road crossing. We present our solution to road crossing challenge for blind individuals. We present the methodology followed in preparing the model and the quantitative assessment of the model. Here first the image is captured through cam which is fixed to forehead and the it sends a model and that transforms into sentence or caption. Here we use API (Google text to voice) its words into voice this voice goes to ear of blind person as an instruction by that blind person can identify easily the things when crossing road

Keywords: Deep Learning, LSTM, Python, VGG16.

I. INTRODUCTION

Many of them with disabilities still find it difficult to fully participate in society, but they are still a valuable and important part of our society. As a result, they have been hampered in their social and economic advancement, and they have little or no desire to contribute to our economic prosperity. Our goal is to assist in bridging this ever-widening gap between the two groups. These technological advancements will assist us in achieving this goal. A person without visual impairments can deduce the scene description and content of an image, but the blind in our society do not have this ability. This ability to provide visual content descriptions in the form of naturally spoken sentences could be extremely beneficial to the visually impaired. If you want to imagine a world where no one is limited by their visual abilities, you can have access to the visual medium without having to see the objects themselves. Their goal is to use an automated method of capturing visual content and producing natural language sentences to empower the visually impaired A new image captioning model known as "domain specific image caption generator" replaces the general caption's specific words with those that are specific to the domain. This model is referred to as a "domain-specific image caption generator" (DSIG). The image caption generator was put to the test in terms of both quality and quantity. The proposed model does not allow for the implementation of a semantic ontology from beginning to end**.**

II. LITERATURE REVIEW

Khadija Essaied et al, Organizational Decision to Adopt Caption Bot Technology-2017 Chatbots have been the subject of several studies over the past decades. The very first chatbot, called Eliza, was developed in 1966 to simulate a consultation with a psychotherapist Since chatbots can ful-fill the role of service employees and are able to support consumers in their decision-making process, they can also act as referral agents or advisors. There is little research in Information Systems (IS) literature on organizational adoption of AI solutions. Some studies investigated the organizational adoption of AI in general, without focusing on any particular type of AI solution. As with most studies on IT organizational adoption, these studies use the Technology Organization Environment (TOE) framework alone or combined with other theories.

Haley Macleod et al, captionbot for assistive vision 2018, the author suggested that BVIPs are trusting of automatically-generated captions, even when they don't make sense. They filled in details or built unsupported narratives to resolve these differences, rather than suspect that captions might be wrong, even when we warned them the captions were authored by a fallible computer algorithm. We conducted an online experiment where we learned that a) our findings from the contextual inquiry study are consistent across a larger sample size and b) negatively framed captions are more appropriate for encouraging appropriate skepticism in uncertain captions (i.e., where congruence and/or confidence are low) understanding blind people's experience with computer-generated Captions of social media images- Edward Cutrell et al,2019 We conducted a small contextual inquiry where we learned that blind and visually impaired people are very trusting of even incorrect AI-generated captions, filling in details to reconcile incongruencies rather than suspecting the caption may be wrong. They described being skilled at detecting incorrect captions and as being consistent about double-checking, but this was not reflected in their behaviour. We then conducted an online experiment to validate these findings on a larger scale. Through this study, we additionally learned that negatively framed captions are best suited to encouraging distrust in incongruent or low confidence captions.

III. PROPOSED METHODOLOGY

A. EXISTING SYSTEM

The existing system was Built using a convolutional neural network (CNN) to extract the visual features, and uses a recurrent neural network (RNN) to translate this data into text. Both CNN and RNN parts can be further trained using the Tensol Flow library However, the existing caption Bot have the disadvantage of They do not make feature observations on objects or actions in the image Information is selected based on the method of maximum sampling It will give up to 60-70% accurate output like object detection in an image.

B. MOTIVATION AND PROBLEM STATEMENT

CNN, LSTM and vocabulary modes are used to caption images. we train LSTM encoder and decoder to identify caption from images and then use vocabulary object to caption/extract meaningful sentence from images or videos. In view of the recent flow of work in intelligent robotics, to our knowledge the results from this research field have scarcely been applied to mitigate sensorial, motor and cognitive damage for humans.

C. PROPOSED SYSTEM

The picture subtitle generator model, we will be blending CNN-RNN structures. Highlight extraction from pictures is finished utilizing CNN. We have utilized the pre-prepared model Exception. The data got from CNN is then utilized by LSTM for creating a depiction of the picture. Nonetheless, sentences that are produced utilizing these methodologies are normally conventional portrayals of the visual substance and foundation data is overlooked. Such conventional portrayals don't fulfil in new circumstances as they, basically repeat the data present in the pictures and nitty gritty depictions with respect to occasions and elements present in the pictures are not given, which is basic to understanding emanant circumstances. Consequently, portraying the substance of a picture is an essential issue in man-made brainpower that interfaces PC vision and regular language preparing. Prior techniques initially produce explanation. The proposed game plan of Image Caption Generator has the capacities to Generate Captions for the Images, given during the Training reason and for the new pictures as well. the Model acknowledges an Image as Input and by inspecting the image it distinguishes objects present in an image and make an engraving which depicts the image okay for any machine to appreciate what an image is endeavouring to state.

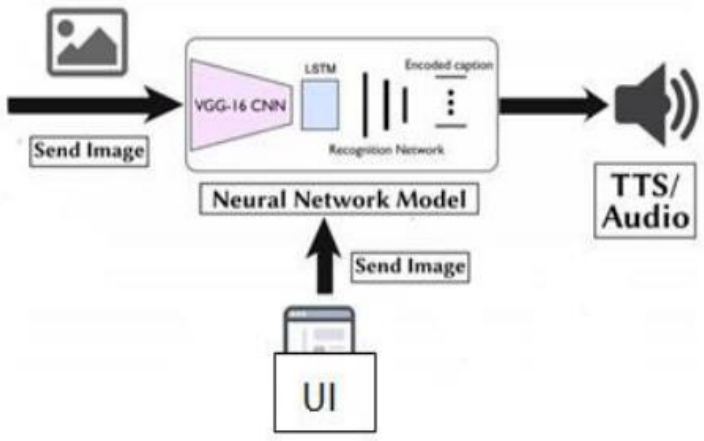

Figure:1 Proposed Architecture

D. MODULES

IMAGE FEATURE EXTRACTION

The best CNN architecture for image classification, the VGG16 model, is used to extract image features. We begin by extracting all of the image's features using this pre-trained model, VGG16. It is possible to save the feature vector created by VGG16. Create an image ID to feature mapping.

TEXT PROCESSING

To begin, lowercase the text and remove all punctuation and word numbers. Create and save a text vocabulary at this point. A mapping between images and descriptions can be created if a single image has multiple descriptions

TOKENIZATION

Beginning and ending symbols should be included in the writing. Tokenizing the text is the final step after adding tokens, and this is where the tokenizer is kept. Tokenize the image's numerical description A sequence of images and words is then used to match the image.

ARCHITECTURE CREATION

Two dense layers represent the image features used for text descriptions. LSTM and embedding are the two methods used. These two networks will be combined to create a network for automatically creating captions for image files

IMAGE CAPTION GENERATION

Creating a human-readable description of a photograph is a difficult artificial intelligence problem. A model from the field of natural language processing is also required for image comprehension. An image can be used to extract features for the pretrained model VGG16. After loading the image, use the saved model and tokenizer to create a caption generation function. Finally, convert the caption into an audio file.

E. IMPLEMENTATION PROCESS

DATASET PREPARATION

The implementation was done by own dataset for this project. We collected around 1000 images of from the internet. After collecting them, we used VGG16 tool to label the images in order to train them after labelling them we saved them in flicker 8 dataset

TRAINING OUR MODEL

The model created a dataset by collecting images with different orientations at different atmospheres from the internet. We downloaded around 1000 images of. After downloading them we need to open caption Bot folder and need to run it then and click generate and load LSTM caption model and need to upload image and extract caption from image then the output will be displayed on the screen with voice and caption in image

F. ALGORITHMS USED

Creating a human-readable description of a photograph is a difficult artificial intelligence problem. A model from the field of natural language processing is also required for image comprehension. An image can be used to extract features for the pretrained model VGG16. After loading the image, use the saved model and tokenizer to create a caption generation function. Finally, convert the caption into an audio file. The best CNN architecture for image classification, the VGG16 model, is used to extract image features. The features of image's were extracted using this pretrained model, VGG16. It is possible to save the feature vector created by VGG16. Create an image ID to feature mapping.

LSTM

LSTM stands for long short-term memory networks, used in the field of Deep Learning. It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems the LSTM is made up of four neural networks and numerous memory blocks known as cells in a chain structure. A conventional LSTM unit consists of a cell, an input gate, an output gate, and a forget gate. The flow of information into and out of the cell is controlled by three gates, and the cell remembers values over arbitrary time intervals. The LSTM algorithm is well adapted to categorize, analyse, and predict time series of uncertain duration.

OPENCV

OpenCV is a computer vision library that contains image-processing algorithms for object detection. OpenCV is a library of python programming language, and real-time computer vision applications can be developed by using the computer vision library. OpenCV library is used in image and video processing and also analysis such as face detection and object detection.

PYTTSX3

pyttsx3 is a text-to-speech conversion library in Python. Unlike alternative libraries, it works offline and is compatible with both Python 2 and 3. An application invokes the pyttsx3.init() factory function to get a reference to a pyttsx3. Engine instance. it is a very easy to use tool which converts the entered text into speech. The pyttsx3 module supports two voices first is female and the second is male which is provided by "sapi5" for window. The text-to-speech features for this module are based on languages installed in your operating system. By default, it should come together with the language pack during the installation of the operating system. You need to install the language pack manually if you intend to use other languages

IV. EXPERMENTAL ANALYSIS

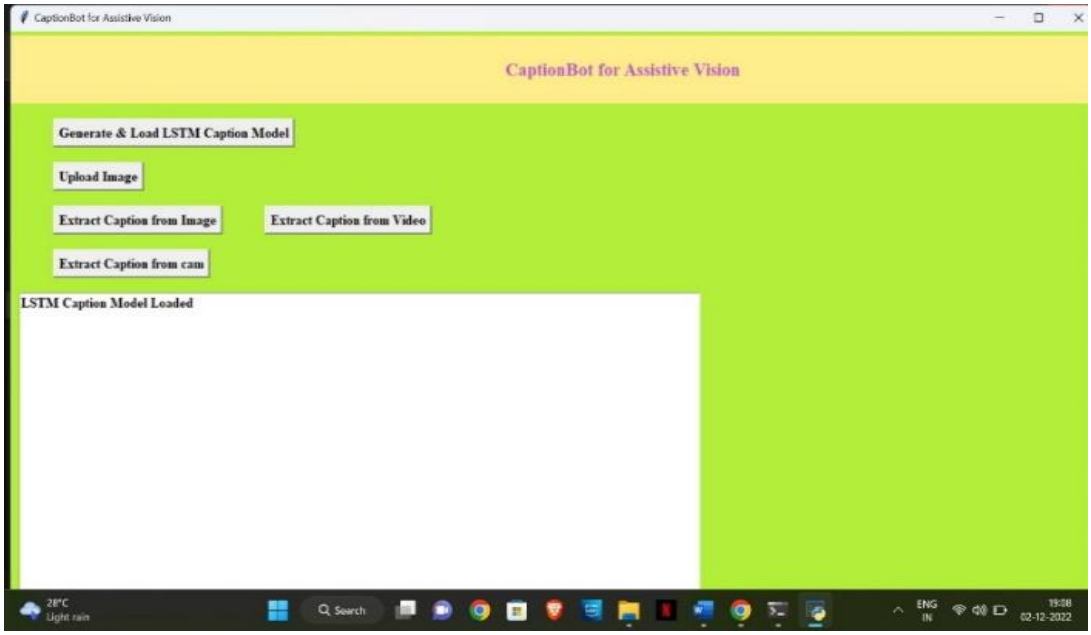

Figure :2 Interface of the Project

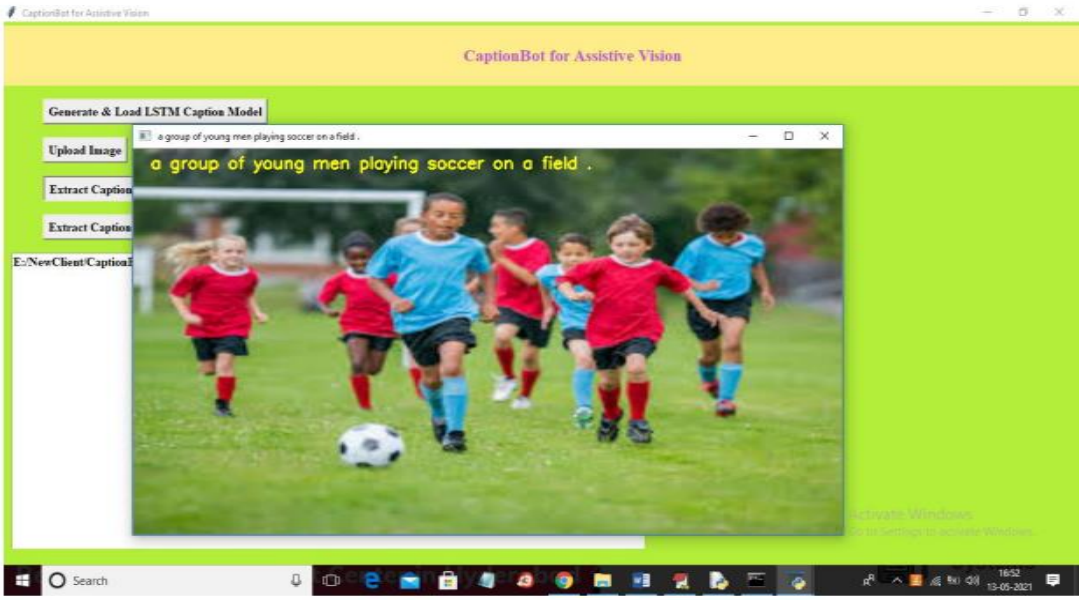

Figure :3 finding the caption from the image

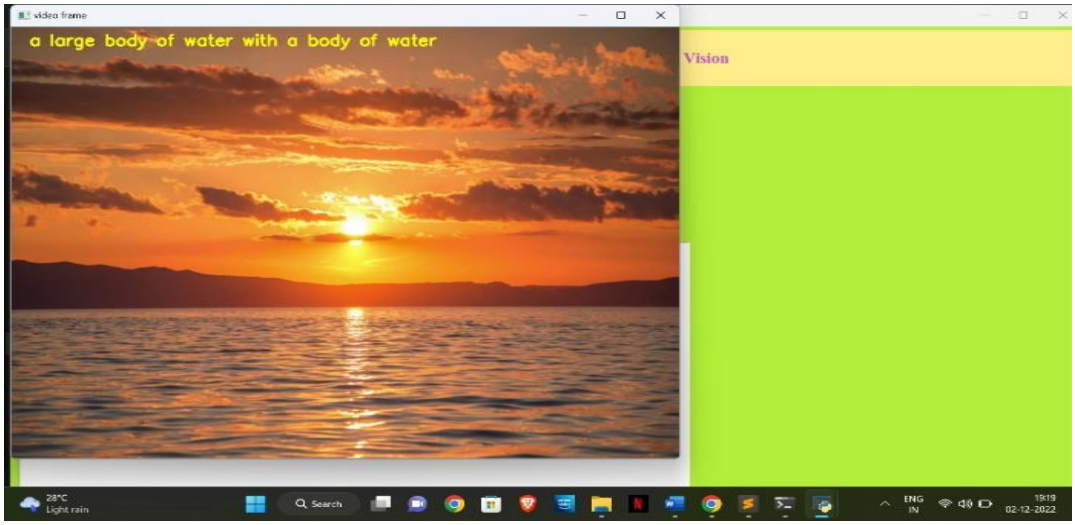

Figure :4 finding the caption from video

V. CONCLUSION

The deep learning model automatically generates image captions with the goal of helping visually impaired people better understand their environments. The described model is based on a CNN that encodes an image into a compact representation, followed by a RNN that generates corresponding sentences based on the learned image features. The proposed model achieved a comparable state-of-the-art performance and that the generated captions are highly descriptive of the objects and scenes depicted on the images. Because of the high quality of the generated image descriptions, visually impaired people can greatly benefit and get a better sense of their surroundings using text-to-speech technology

Future work can include this text-to-speech technology, so that the generated descriptions are automatically read out loud to visually impaired people. In addition, future work could focus on translating videos directly to sentences instead of generating captions of images. Static images can only provide blind people with information about one specific instant of time, while video caption generation could potentially provide blind people with continuous real time information. LSTMs could be used in combination with CNNS to translate videos to English descriptions.

VI. REFERENCES

[1] Yimin Zhou, Yiwei Sun, Vasant Honavar, (2019) s" Improving Image Captioning by Leveraging Knowledge Graphs" IEEE pp 99-103
[2] Krizhevsky, I. Sutskever, and G. E. Hinton, (2012) "Imagenet classification with deep convolutional neural networks," in NIPS'12, vol. 1, pp. 1097–1105
[3] R. Krishna et al., (2017) "Visual genome: Connecting language and vision using crowdsourced dense image annotations," International Journal of Computer Vision, vol. 123, no. 1, pp. 32–73
[4] G. Kulkarni et al. (2013), "Babytalk: Understanding and generating simple image descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891–2903.
[5] Itunuoluwa Isewon , Jelili Oyelade ,Olufunke Oladipupo(2014) "Design and Implementationof Text To Speech Conversion for VisuallyImpaired People" , International Journal ofApplied Information Systems (IJAIS) – ISSN :2249-0868 Foundation of Computer Science FCS,New York, USA Volume 7– No. 2

[6] Ignatious, L.A.A., Jeevitha, S., Madhurambigai, M. and Hemalatha, M., (2019), December. A Semantic Driven CNN–LSTM Architecture for Personalised Image Caption Generation. In 11th International Conference on Advanced Computing (ICoAC) pp. 356-362.

[7] Wang, S., Tian(2019), Y: Camera-based signage detection and recognition for blind persons. In: Miesenberger, K., Karshmer, A., Penaz, P., Zagler, W. (eds.) Computers Helping People with Special Needs, pp. 17–24

[8] Sanjana, B., RejinaParvin, J(2016).: Voice assisted text reading system for visually impaired persons using TTS method, 2016. IOSR J. VLSI Sig. Process. **6**(3), Ver. III, pp. 15–23

[9] Wang, R. and Wakahara, T., (2019), July. Practice in Caption Generation with Keras: The Design and Evaluation for Attention Models. In Proceedings of the 2019 3rd International Conference on Deep Learning Technologies pp. 11-15.

[10] Iyer, S., Chaturvedi, S. and Dash, T., (2019). Image captioning-based image search engine: An alternative to retrieval by metadata. In Soft Computing for Problem Solving pp. 181-191.

[11] Amirian, S., Rasheed, K., Taha, T.R. and Arabnia, H.R., (2019). A short review on image caption generation with deep learning. In Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV) pp. 10-18.

[12] Shetty, S., Hegde, S., Shetty, S., Shetty, D., Sowmya, M.R., Miranda, R., Sequeira, F. and Menezes, J., (2022). Deep Learning Photograph Caption Generator. In Recent Advances in Artificial Intelligence and Data Engineering pp. 277-288.

[13] Srivastava, S., Sharma, H. and Dixit, P., (2022), January. Image Captioning based on Deep Convolutional Neural Networks and LSTM. In 2022 2nd International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC) pp. 1-4.