Image Descriptor For Visually Impaired

1st Athulya Anilkumar
Dept. Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
athulyaanilkumar0709@gmail.com

2nd Abhinav V V
Dept. Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
abhinavvv18@gmail.com

3rd Aneeta Shajan
Dept. Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
aneetashajan999@gmail.com

4th Anjana S Nair
Dept. Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
anjananair3030@gmail.com

5th Bini M Issac
Dept. Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
binimissac@amaljyothi.ac.in

6th Neenu R
Dept. Computer Science & Engineering
Amal Jyothi College of Engineering
Kottayam, India
rneenu@amaljyothi.ac.in

***Abstract*—An image descriptor is a system that generates a voice description of the context of an image. The initial step involves the generation of a textual description of the image. It entails analyzing an image with machine learning algorithms and producing a description of the image in natural language. The obtained captions are then converted to voice output. Systems for captioning images can be used for a variety of purposes, including assisting those who are visually impaired in comprehending what is being depicted in a picture or assisting search engines in comprehending picture content and enhancing search results. Building systems for captioning images can be done in a number of ways. One method entails employing a neural network to compress the image into a representation, followed by another neural network to decode the representation into a description in natural language. Typically, a large collection of photos and videos is used to train the neural networks.**

I. INTRODUCTION

Image captioning is an active area of research in the field of artificial intelligence and machine learning. The development of image captioning systems has been influenced by several factors, including advances in deep learning and natural language processing, as well as the availability of large datasets of images and text descriptions. One of the earliest approaches to image captioning involved using hand-crafted features to represent the content of an image and then using these features to generate a textual description. However, these systems were limited in their ability to accurately describe complex images. With the advent of deep learning, researchers were able to develop more powerful image captioning systems that could analyze images at a higher level of abstraction. These systems used neural networks to learn a compact representation of the image and then used a separate neural network to decode the representation into a natural language description. In recent years, there have been many significant advances in image captioning, including the development of more powerful neural network architectures, the use of attention mechanisms to focus on specific regions of the image, and the use of large-scale datasets to train image captioning systems.

II. LITERATURE SURVEY

A model for image captioning in the Hindi language is developed. The dataset is manually created by translating well known MSCOCO dataset from English to Hindi. Also, different types of attention-based architectures are developed for image captioning in the Hindi language. The obtained results of the proposed model are compared with several baselines in terms of BLEU scores [1].

A review of different architectures used for image captioning using different datasets and evaluated using different evaluation metrics.By this comparison, it is clear that Anderson et al. performed well on the MSCOCO dataset. Their method has surpassed previous work. This is because it uses an attention mechanism that focuses only on relevant objects in the image [5].

LRCN, a class of models that is both spatially and temporally deep, and flexible enough to be applied to a variety of vision tasks involving sequential inputs and outputs is introduced. Results show that such models have distinct advantages over models for recognition or generation which are separately defined or optimized [7].

A survey that provides an overview of image captioning methods, from technical architectures to datasets, evaluation metrics, and a comparison of recent approaches.Aims to discover an efficient method for processing the query image, representing its content, and transforming it into a sentence by creating connections between textual and visual elements while maintaining language fluency [8].

An introduction of algorithms and techniques used in the field of text generation like distributed representation of words CNN, LSTM, BRNN etc. Also, the activation functions like the Sigmiod, and the optimization techniques like the Stochastic Gradient Descent (SGD) are mentioned along with the recent techniques for text generation like VAE and GAN [9].

A discussion on the classification of various image classifiers used for image classification where its fundamental concepts are mainly discussed along with their advantages

and disadvantages for the following like KNN classifier,Linear classifier,Softmax,CNN and arrived the conclusion that among these CNN shows the best results. [12].

A deep learning method for generating captions with object detection and feature extraction using neural networks includes various steps such as object detection,creating attributes,encoder and decoder.The method was tested on a Flickr 8k dataset. Compared to the already available picture caption creation generators, the suggested deep learning methodology produced captions with greater descriptive meaning. [11].

III. METHODOLOGY

The basic steps involved in building our model are explained below with the help of a figure 1:

- Collect a dataset of images and their corresponding captions. This dataset will be used to train the image captioning model. Here the dataset we have decided is Flickr 30k. It is a dataset used for understanding the visual media that correspond to a linguistic expression and it contains 31,000 images collected from Flickr, together with 5 reference sentences provided by human annotators.
- The model choosen here is a combination of Convolutional Neural Network(CNN) and Bidirectional Long Short Term(Bi-LSTM).
- Train the model on the dataset. This will involve feeding the images and captions into the model, and using an optimizer to adjust the model's weights and biases to minimize the loss function. Optimizers like the Stochastic Gradient Descent can be used.
- Evaluate the model on a separate test set. This will help to determine how well the model is able to generate accurate captions for images it has not seen before.
- Fine-tune the model. Adjusting the model architecture, training on a larger dataset, or using different optimization techniques to improve the model's performance.
- Test the trained model on a separate dataset to evaluate its performance. This can be done using various metrics, such as BLEU scores or METEOR scores, which measure the similarity between the generated text and a reference description.

Once the model is created, Google text to speech API is used for generating voice output for the generated captions. The user interface for android app is created using Kivy framework in python.

A. Architecture

After performing the extensive literature survey, it was found that the most widely used approach involves using a neural network to encode the image into a compact representation, and then using a separate neural network to decode the representation into a natural language description. The neural networks are typically trained on a large dataset of images and their corresponding text descriptions, and the goal is to learn a model that can generate accurate descriptions for new images. The figure 2 shows the architecture of our image captioning model.

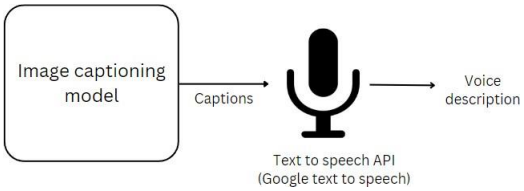

Fig. 1. Basic architecture of proposed system

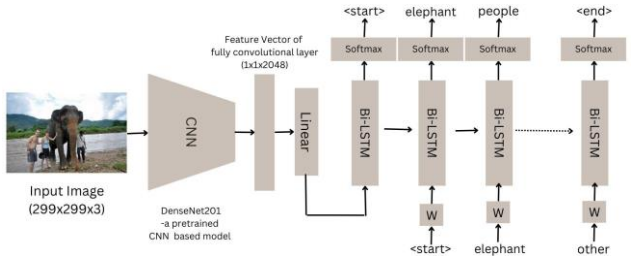

Fig. 2. Architecture of Image Captioning Model

Convolutional Neural networks-Convolutional Neural networks are specialized deep neural networks which is able to process the data that has input shape like a 2D matrix. Images can easily be represented as a 2D matrix and CNN is very useful in working with images. CNN is basically used for image classifications and identifying if an image is a tiger, a car or a plant etc. It scans images from left to right and top to bottom to find important features from the image and combines the feature to classify images. It can handle the images that have been translated, rotated, scaled and changes in perspective. We are using ResNet101 which is a pretrained CNN based model as the encoder.

Bi-directional Long Short-Term Memory-Bi-LSTM networks can be used in the process of image captioning, which involves generating a natural language description of an image. In an image captioning system that uses a Bi-LSTM, the Bi-LSTM is typically used to decode the compact representation of the image, which is generated by a convolutional neural network (CNN). The CNN encodes the image into a fixed-length vector, which is then input to the Bi-LSTM. The Bi-LSTM processes the vector and generates a sequence of words that form the caption for the image. One advantage of using a Bi-LSTM in image captioning is that it can consider both past and future context when

generating the caption. This can be important because the words in a caption are often interdependent, and the meaning of a word can depend on the words that come before and after it. There are several ways to train a Bi-LSTM for image captioning, such as using supervised learning with a large dataset of images and their corresponding text descriptions, or using reinforcement learning to optimize the Bi-LSTM for a specific task or metric.

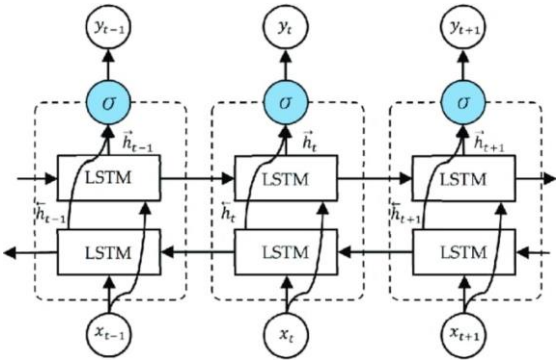

Fig. 3. Bi-LSTM architecture

B. Technologies used

Google Cloud Text-to-Speech- Google Cloud Text-to-Speech is a cloud-based service that converts text into natural-sounding synthesized speech. The service uses machine learning algorithms to synthesize speech that is similar to human voice patterns, making it possible to generate speech that is natural and easy to understand. One of the key features of Google Cloud Text-to-Speech is its support for a wide range of languages and voices. The service currently supports over 180 voices in 30 languages, including Malayalam. This makes it possible to use the service to generate speech in a variety of languages and accents, making it suitable for a wide range of use cases.

TensorFlow- TensorFlow is a free and open-source software library. It can be used to solve a variety of problems, but the focus is on training and inference of deep neural networks. It is a symbolic math library used for machine learning applications such as neural networks. It allows developers to create data flow graphs to easily build and deploy machine learning models. In image captioning models, TensorFlow is used to build and train deep neural networks that can understand natural language and interpret images. By training the model with a large set of images and captions, the model can learn to map an image to a caption. The model can then be used to generate captions for new images. TensorFlow can also be used to optimize the model parameters such as the learning rate, number of layers, and number of neurons. This makes the model more accurate and faster.

Kivy- Kivy is a free and open source Python framework for developing mobile apps and other multitouch application software with a natural user interface (NUI). It is designed to be highly flexible and customizable, allowing developers to create applications for a wide range of platforms. Kivy can be used in image captioning models by providing the user interface and user experience. Kivy allows developers to create a user interface for an image captioning model quickly, with features such as drag and drop, custom UI elements, and live previews. This makes it easy and fast for users to interact with the model and generate captions for images. Kivy also allows developers to create custom UI elements such as buttons, sliders, and text fields that can be used to customize and control the image captioning model. Furthermore, Kivy can be used to create a mobile app for the model, allowing users to access the model on their mobile devices.

C. Dataset preparation

The Flickr30k dataset is a collection of images and their corresponding text descriptions. It consists of approximately 30,000 images, each with five different text descriptions written by different annotators. The dataset is often used to evaluate and compare the performance of image captioning systems.

This dataset is widely used for the purpose of image captioning and it does not require any pre processing steps before using it for training the model. Flickr30k Data Fields:

- images: tensor containing the image
- texts: tensor to represent text associated with the image.
- comment_nos: tensor to represent the number of comments.

IV. EVALUATION

The model is expected to have lesser errors and also reduced loss function than the existing models. The model is trained to generate captions that resonate well with visually impaired people.

A. Comparison of a CNN and LSTM model with our CNN and Bi-LSTM model

A basic CNN-LSTM image caption generating model typically consists of a convolutional neural network (CNN) that processes the input image and extracts its features, followed by a long short-term memory (LSTM) network that generates the corresponding caption. The CNN extracts spatial features from the image, while the LSTM is responsible for generating the caption by taking into account the temporal dependencies between the words.

On the other hand, this CNN-BiLSTM image caption generating model adds a bidirectional LSTM (BiLSTM) to the CNN-LSTM architecture. The BiLSTM network processes the output of the CNN and generates a sequence of feature vectors that contain information about the input image in both forward and backward directions. This allows the model to better capture the dependencies between the image and the generated caption, leading to more accurate and descriptive

captions.

In terms of performance, the CNN-BiLSTM image caption generating model generally outperforms a CNN-LSTM model due to its ability to capture more complex dependencies between the image and the generated caption. The bidirectional nature of the BiLSTM allows the model to better understand the context and meaning of the words in the caption, leading to more fluent and natural-sounding captions.It can perform better due to its ability to capture more complex dependencies and generate more natural-sounding captions.

B. Evaluation metric

BLEU (Bilingual Evaluation Understudy), is a metric for evaluating the quality of machine-generated translations, especially in the context of natural language processing (NLP) tasks. BLEU score measures the similarity between the machine-generated translation and one or more human reference translations. It does this by computing the n-gram overlap between the machine-generated output and the reference translations.The score is based on a weighted geometric mean of n-gram precisions, where n is typically 1, 2, 3, or 4. The weights are based on the lengths of the n-grams, with shorter n-grams being given more weight. The final score ranges from 0 to 1, with higher scores indicating better translation quality.

Our model has a BLEU score of 0.64. A Bleu score of 0.64 indicates that the generated captions are moderately similar to the human-written captions for the given images.The score suggests that the image caption generator has achieved a reasonable level of accuracy in generating captions that are comparable to human-generated captions.

V. CONCLUSION

The development of an effective image descriptor for visually impaired individuals is a crucial step toward improving their ability to understand and interact with the visual world. In this project, we aim to explore the use of deep learning methods, including convolutional neural networks (CNN) and bidirectional long short-term memory (BiLSTM) networks, to create such a descriptor. For this, we are planning to use the Flickr30K dataset, which consists of a large collection of images annotated with descriptive captions.CNN will be used as an encoder to extract features from the images and BiLSTM will be used as a decoder to generate descriptive captions based on these features. The CNN-BiLSTM image descriptor is expected to achieve high accuracy levels on various image classification tasks. This project has the potential to significantly improve the ability of visually impaired individuals to understand and interact with the visual world. Further research could be conducted to improve the performance of the model, such as exploring different CNN and BiLSTM architectures or incorporating additional datasets. Additionally, future work could also focus on integrating the image descriptor into a larger assistive technology system for visually impaired individuals, such as a screen reader or mobile application.

References

[1] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha and Pushpak Bhat- tacharyya, “A Hindi Image Caption Generation Framework Using Deep Learning” ACM Transactions on Asian and Low-Resource Language Information ProcessingVolume 20Issue 2March 2021 Article No.: 32pp 1–19https://doi.org/10.1145/3432246

[2] M. P. R, M. Anu and D. S, “Building A Voice Based Image Caption Generator with Deep Learning,” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 943- 948, doi: 10.1109/ICICCS51141.2021.9432091.

[3] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542, 1 July 2022, doi: 10.1109/TPAMI.2021.3059968.

[4] Ariyo Oluwasammi, Muhammad Umar Aftab, Zhiguang Qin, Son Tung Ngo, Thang Van Doan, Son Ba Nguyen, Son Hoang Nguyen, Giang Hoang Nguyen, “Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Caption- ing”, Complexity, vol. 2021, Article ID 5538927, 19 pages, 2021. https://doi.org/10.1155/2021/5538927

[5] S. Amirian, K. Rasheed, T. R. Taha and H. R. Arabnia, “Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap,” in IEEE Access, vol. 8, pp. 218386- 218400, 2020, doi: 10.1109/ACCESS.2020.3042484.

[6] G. Cheng, C. Ma, P. Zhou, X. Yao and J. Han, “Scene classi- fication of high resolution remote sensing images using convolu- tional neural networks,” 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2016, pp. 767-770, doi: 10.1109/IGARSS.2016.7729193.

[7] J. Donahue et al.,“Long-Term Recurrent Convolutional Networks for Visual Recognition and Description,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 677-691, 1 April 2017, doi: 10.1109/TPAMI.2016.2599174.

[8] Luo, Gaifang & Cheng, Lijun & Jing, Chao & Zhao, Can & Song, Guozhu. (2022). A thorough review of models, evaluation metrics, and datasets on image captioning. IET Image Processing. 16.10.1049/ipr2.12367.

[9] Touseef Iqbal, Shaima Qureshi, The survey: Text generation models in deep learning, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 6, Part A, 2022, Pages 2515- 2528, ISSN 1319-1578, https://doi.org/10.1016/j.jksuci.2020.04.001.

[10] Boyko, Nataliya, Iryna Khomyshyn, Natalia Ortynska, Viktoria Terlet- ska, Oksana Bilyk and Oleksandra Hasko. “Analysis of the Application of Stochastic Gradient Descent to Detect Network Violations.” COLINS(2022).

[11] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman and J. Yuvaraj, “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach,” 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, pp. 107-109, doi: 10.1109/ICACCS.2019.8728516.

[12] Jogin, Manjunath & Mohana, Mohana & Madhulika, M & Divya, G & Meghana, R & Apoorva, S. (2018).

Feature Extraction using Convolution Neural Networks (CNN) and Deep Learning. 2319-2323. 10.1109/RTEICT42901.2018.9012507.

[13] Selvaraj, Chithra & Natarajan, Bhalaji. (2018). Enhanced portable text to speech converter for visually impaired. International Jour- nal of Intelligent Systems Technologies and Applications. 17. 42. 10.1504/IJISTA.2018.10012881.