AI Based Image Caption Generator with Aid for Visually and Verbally Impaired People

Mandar Hadap[1], Shivam Singh[2], Shubham Palekar[3], Katherine Nadar [4]

Prof. Deepa Athawale[5]

[1]B.E Student of Bharat College Of Engineering, Badlapur
[2]B.E Student of Bharat College Of Engineering, Badlapur
[3]B.E Student of Bharat College Of Engineering, Badlapur
[4]B.E Student of Bharat College Of Engineering, Badlapur
[5]B.E Professor of Bharat College Of Engineering, Badlapur

Abstract - This AI based image caption generator is developed for the purpose for image caption generating so that it helps the visually and verbally impaired individuals. The objective of this is that a Blind person can upload the image and can get a description of the image that can help them understand it. Here text to speech module is also used so that it can be heard by the blind individual.

***Key Words*:** Artificial Intelligence, Machine learning, Captioning, LSTM (Long Short-Term Memory, RNN (Recurrent Neural Network), CNN (Concurrent Neural Network), CV (Computer Vision). insert (key words)

1.INTRODUCTION

In today's visually-driven digital world, images are a primary medium of communication and expression. Social media platforms, educational content, and online services heavily rely on visual content to engage users. However, this reliance on images can pose significant challenges for visually and verbally impaired individuals who cannot easily access or interpret such content. The AI-Based Image Caption Generator is a web application developed to enhance accessibility for visually and verbally impaired individuals. In a world increasingly dominated by visual content, many people face barriers to accessing and understanding images shared across platforms like social media, educational websites, and digital communication tools. This project aims to address these challenges by leveraging artificial intelligence to generate descriptive captions for images automatically.

2. PROBLEM STATEMENT

Current image captioning systems lack comprehensive accessibility features, making it difficult for visually and verbally impaired individuals to fully engage with visual content. Existing solutions often separate image captioning from text-to-speech and translation services, creating barriers to seamless understanding and interaction. There is a need for an integrated platform that provides accurate image descriptions, supports multiple languages, and offers intuitive accessibility features within a single, scalable application.

Existing image captioning systems often fail to address the needs of users with disabilities effectively. While some offer image descriptions, they typically do not integrate text-to-speech or translation functionalities, limiting accessibility. Users with visual or verbal impairments may struggle to understand and interact with visual content fully. Therefore, there is a need for a unified solution that combines automatic image captioning, multilingual support, and text-to-speech capabilities to enhance accessibility and user experience for all.

3. SYSTEM METHODOLOGY

A. Convolutional Neural Network (CNN)

The CNN is designed in such a way that the benefit of 2D structure of input image can be taken. This target is accomplished with the help of number of local connections and tied weights along with various pooling techniques which result in translation invariant features. The convolutional networks are currently used in visual recognition. There are number of convolutional layers in CNN. After these convolutional layers, next layers are fully connected layers as in multilayer neural network [14]. The main advantages of using CNN are ease of training and possessing less parameters as compared to

other networks with equal number of hidden states. For this work, we are using Visual Group Geometry (VGG) network, which is Deep CNN for large scale image recognition [15]. It is available in 16 layers as well as 19 layers. The classification error results for both 16 and 19 layers are almost same for validation set as well as test set, which is around 7.4% and 7.3%. This model gives the features of images which are used in further process of caption generation.

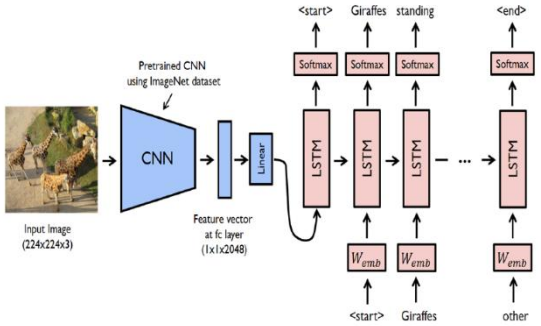

Fig -1: Convolutional Neural Network (CNN)

B. Recurrent Neural Network (RNN)

Recurrent Neural Network (RNN) is a type of neural network designed for sequential data, such as text, speech, and time series. Unlike traditional neural networks that process inputs independently, RNNs have loops that allow them to retain information from previous inputs, making them useful for tasks where context is important.

How Does RNN Work?

Instead of treating each input independently, an RNN **remembers previous inputs** through **hidden states**. It processes sequences step by step while retaining relevant past information.

- **Input Layer** → Takes in sequential data (e.g., a sentence: "The cat is sleeping").
- **Hidden Layer** → Processes each word and maintains memory through recurrent connections.
- **Output Layer** → Generates predictions (e.g., next word in a sentence).
-

At each time step **t**, the RNN updates its hidden state:

$h_t = f(W \cdot x_t + U \cdot h_{t-1} + b)$……(i)

where:

- x_t is the input at time t
- h_{t-1} is the previous hidden state
- W, U are weight matrices
- b is the bias
- f is an activation function (like Tanh or ReLU)

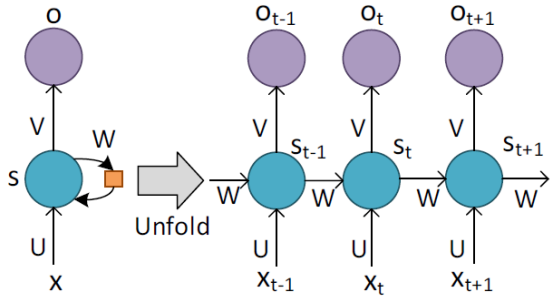

Figure 2. The model structure of RNN

C. Long Short-term Memory (LSTM)

The transitory dynamics in a set of things are modelled by using a recurrent neural network [17]. It is very difficult for ordinary RNN to acquire long term dynamics as they get vanished and exploding weights or gradients [9]. The memory cell is main block of LSTM. It stores the present value for long period of time. Gates are there for controlling update time of state of cell. The number of connections between memory cell and gates represent variants.

Our model is based on the LSTM block which depends on the LSTM with no peephole architecture as shown in Fig. 3. The memory cell and gates of LSTM are having following relations:

$$i_l = \sigma(W_{ix} x_l + W_{im} m_{l-1}) \quad (1)$$

$$f_l = \sigma(W_{fx} x_l + W_{fm} m_{l-1}) \quad (2)$$

$$o_l = \sigma(W_{ox} x_l + W_{om} m_{l-1}) \quad (3)$$

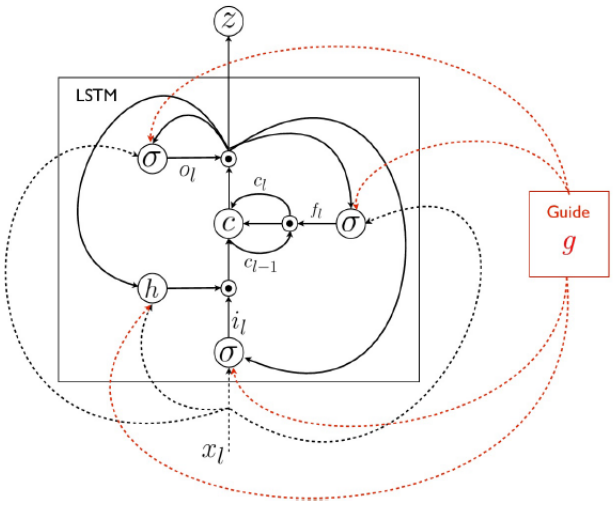

Fig. 3. Connection diagram of LSTM [9]

$$c_l = f_l \odot c_{l-1} + i_l \odot \hbar(W_{cx} x_l + W_{cm} m_{l-1}) \quad (4)$$

$$m_l = o_l \odot c_l \quad (5)$$

$$L(I,S) = -\sum_{t=1}^{N} log(p_t(S_t)) \quad (6)$$

D. Image Captioning and text Generation

Image caption is a basic multimodal problem in the field of artificial intelligence, which connects computer vision with natural language generation. It can be divided into two steps, feature extraction and natural language generation. Kiro set al. [32] introduced the neural language model of multimodal
constraint and used CNN to learn the word representation and
image features together. Vinyals et al. [33] proposed a generation model based on deep RNN architecture. Given the training image, the model could be trained to maximize the probability of the target sentence

(i) BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

Vision-Language Pre-training (VLP) has advanced the performance for many vision-language tasks. However, most existing pre-trained models only excel in either understanding-based tasks or generation-based tasks. Furthermore, performance improvement has been largely achieved by scaling up the dataset with noisy image-text pairs collected from the web, which is a suboptimal source of supervision. In this paper, we propose BLIP, a new VLP framework which transfers flexibly to both vision-language understanding and generation tasks. BLIP effectively utilizes the noisy web data by bootstrapping the captions, where a captioner generates synthetic captions and a filter removes the noisy ones. We achieve state-of-the-art results on a wide range of vision-language tasks, such as image-text retrieval (+2.7% in average recall@1), image captioning (+2.8% in CIDEr), and VQA (+1.6% in VQA score). BLIP also demonstrates strong generalization ability when directly transferred to video language tasks in a zero-shot manner. Code, models, and datasets are released.

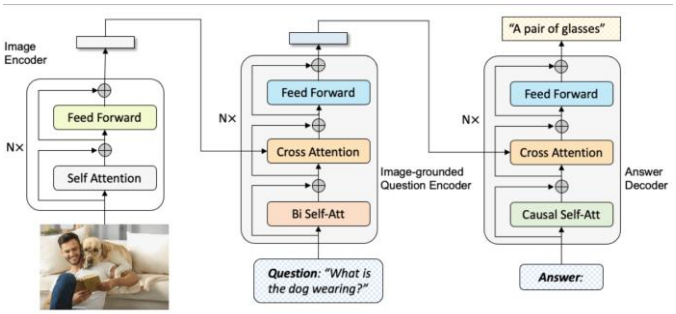

Figure 4. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

(i) SpeechT5 (TTS task)

T5 (Text-To-Text Transfer Transformer) in pre-trained natural language processing models, we propose a unified-modal SpeechT5 framework that explores the encoder-decoder pre-training for self-supervised speech/text representation learning. The SpeechT5 framework consists of a shared encoder-decoder network and six modal-specific (speech/text) pre/post-nets. After preprocessing the input speech/text through the pre-nets, the shared encoder-decoder network models the sequence-to-sequence transformation, and then the post-nets generate the output in the speech/text modality based on the output of the decoder.

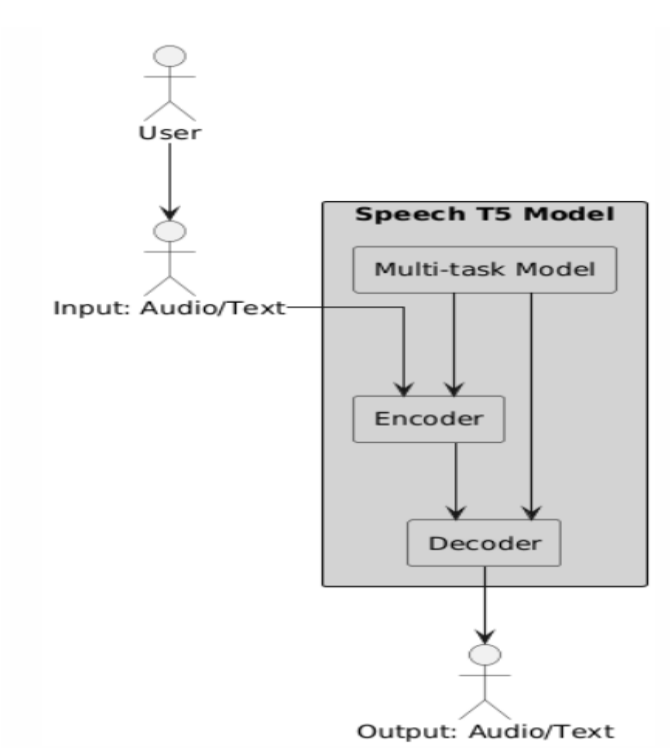

Figure 2. Speech T5 Architecture

E. Results

(i) Datasets

These datasets consist of images and description of image in the form of sentences in natural language such as English. The statistics of datasets are as shown in Table I. In these datasets, each image is described by observers with 5 different sentences that are relatively visible and impartial.

TABLE I
DATASET STATISTICS

Dataset Name	Size		
	Train	*Valid*	*Test*
Flickr8k [1]	6000	1000	1000
Flickr30k [1]	28000	1000	1000
MSCOCO [1]	82783	40504	40775

(ii) Results

The model has been trained for 50 epochs. As number of epochs used are more, it helps to lower the loss to 3.74. If we consider the large dataset then we should use more epochs for accurate results. Some results generated are as shown in Fig. 4. By using the Flickr8k dataset for training model and running test on the 1000 test images available in dataset results in BLEU = 0.53356. For Flickr30k dataset, running test on same number of test images available in dataset results in BLEU = 0.61433 and for MSCOCO dataset running test on images results in BLEU = 0.67257.

F. Conclusion

The AI-Based Image Caption Generator project successfully demonstrates the potential of leveraging AI to improve accessibility for visually and verbally impaired individuals. By incorporating advanced techniques such as convolutional neural networks (CNN) and natural language processing (NLP), the application efficiently generates accurate and descriptive captions for images. The integration of features like text-to-speech and Braille script enhances usability for a diverse audience. Through continuous refinement and performance optimization, the system has proven effective in generating meaningful image descriptions.

- The project successfully integrates advanced AI techniques to enhance accessibility for impaired users.
- It demonstrates the practical application of AI in real-world scenarios, particularly for assistive technology.
- The robust and user-friendly interface, powered by the MERN stack, ensures seamless interaction and engagement.

REFERENCES

1. N. Komal Kumar1, D. Vigneswari2, A. Mohan3, K. Laxman4, J. Yuvaraj5: Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS)
2. Qiuyun Zhang, Bin Guo, Hao Wang, Yunji Liang, Shaoyang Hao, Zhiwen Yu.: AI-Powered Text Generation for Harmonious Human-Machine Interaction: School of Computer Science, Northwestern Polytechnical University, Xi'an 710129, P.R.China

guob@nwpu.edu.cn:2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation

3. Sruthi K V, Meharban M S: Department of Computer Science and Engineering Rajagiri School of Engineering and Technology Ernakulam, India, 2020 6th International Conference on Advanced Computing & Communication Systems (ICACCS)
4. Chetan Amritkar, Vaishali Jabade .: Department of EnTC Vishwakarma Institute of Technology Pune, India chetan.amritkar16@vit.edu. 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)
5. Ponnaganti Rama Devi1, Mannam Thrushanth Deepak2, Morampudi Lohitha3, M.Surya Chandra Raju 4, K.Venkata Ramana5 2,3,4,5 Student, Department of computer science engineering Gitam, Visakhapatnam, Andhra Pradesh, India 1 Assistant professor, Department of computer science engineering Gitam, Visakhapatnam, Andhra Pradesh, India International Journal of Advances in Engineering and Management (IJAEM) Volume 5, Issue 4 April 2023, pp: 576-583

BIOGRAPHIES

Mandar Vijay Hadap

Shivam Sujit Singh

Shubham Mahendra Palekar

Katherine Nadar