

Enhancing accessibility: deep learning-based image description for individuals with visual impairments

Nidhi B. Shah¹, Amit P. Ganatra²

¹Computer Engineering Department, Charusat University, Gujarat, India

²Parul University, Vadodara, India

Article Info

Article history:

Received Jul 24, 2024

Revised Nov 4, 2024

Accepted Nov 11, 2024

Keywords:

CNN

Deep learning

Image processing

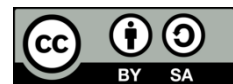
LSTM

RNN

ABSTRACT

Technological developments in artificial intelligence, namely in the area of deep learning, have created new avenues for enhancing accessibility for those with visual impairments. In order to improve the capacity of people who are blind or visually impaired to understand and interact with visual material, this research investigates the creation and use of deep learning-based image description systems. We provide a comprehensive method that uses recurrent neural networks (RNNs) to generate natural language descriptions and convolutional neural networks (CNNs) and Autoencoders for extracting picture features. Our technology automatically creates comprehensive, context-aware descriptions of photographs by incorporating these models, giving users a better knowledge of their surroundings. We show the accuracy and reliability of the system on a wide range of photos through comprehensive testing. According to our research, deep learning-based picture description systems and converting the description in audio and making a promise to empower people who are visually impaired and foster diversity in the digital sphere.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Nidhi B. Shah

Computer Engineering Department, Charusat University

India

Email: nbshah999@yahoo.com

1. INTRODUCTION

Typically, encoder-decoder architectures are used for image caption models, where captions are generated by feeding abstract picture feature vectors into an encoder. An important problem in computer vision, natural language processing, artificial intelligence, and image processing is producing natural language descriptions from images. Scene understanding, which integrates computer vision and natural language processing expertise, includes image captioning, which automatically produces natural language descriptions based on what is observed in the image.

Existing algorithms, which are a combination of convolutional and recurrent networks used to generate captions, have many problems, such as missing gradients, not accurately identifying objects and relationships, or generating captions only for visible images. Model for automatic generation of captioned images by combining advanced convolutional memory deep neural networks (CNN) and long-short-term (LSTM) memory and algorithms. This is a variant on the conventional approach meant to address issues that crop up when employing conventional subtitling techniques. There are two steps to the model. Convolutional algorithms are used in the first stage, while long-term memory is used in the second. Image is the input for the first phase. Additionally, the useful captions that effectively depict the visual scene are a key component of the suggested system model.

To develop a system that generates accurate signatures from input images using CNN and LSTM algorithms. CNNs are used to extract features from image. A CNN is a specialized deep neural network that can process data with an input shape such as a two-dimensional matrix. Images are easy to represent as a 2D matrix, and CNNs are very useful for working with images. CNNs are mainly used to classify images and determine whether the image is a bird, an airplane, and superman. It extracts significant information from the image by scanning it from top to bottom and left to right, then combines these features to categorize the image. It can manage perspective-adjusted, rotated, scaled, and translated images. A basic captioning system flow diagram is Figure 1.

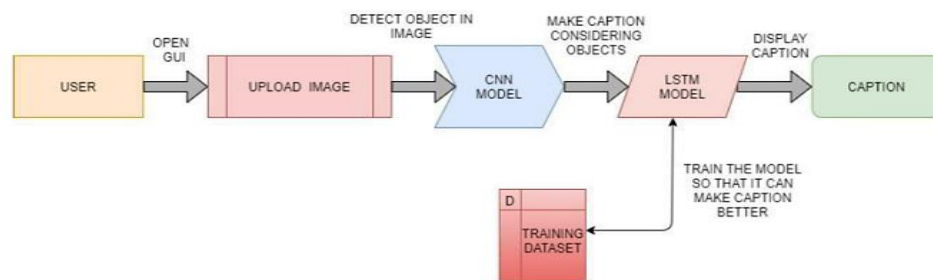


Figure 1. Simple flow diagram of system

LSTM uses CNN information to generate a description of an image. LSTM is a type of recurrent neural network (RNN) that is very suitable for sequence prediction tasks. Based on past text, it can anticipate the word that will appear next. By getting beyond the drawbacks of existing RNNs with short-term memory, it was demonstrated to be more effective than those RNNs. With the aid of a forget gate, an LSTM may retain pertinent information.

2. RELATED WORK

Visual accessibility is a critical issue that affects millions of individuals worldwide. To address this challenge, researchers have explored the potential of deep learning-based image description systems to provide detailed, contextual information about visual content to those with visual impairments [1]. Implementation and evaluation of autoencoders are done. Training begins with a raw picture. Image prediction using training data is being tested. Precision and time are measured in this investigation. The model may also be used to input an encoded picture and a decoded image, the autoencoder's intermediate and final outputs. Different autoencoders were used to compare model outcomes [2], [3].

Recent advancements in computer vision and natural language processing have led to the development of AI-driven image captioning tools that can automatically generate textual descriptions of visual content. These systems have the potential to empower individuals with visual impairments by allowing them to access and comprehend the visual information that is increasingly prevalent in digital media and social platforms [4]. However, the effectiveness of these tools is contingent on their ability to generate descriptions that are not only accurate but also tailored to the specific needs and preferences of visually impaired users.

Existing image captioning systems have been primarily evaluated based on their similarity to human-authored descriptions, which may not necessarily align with the informational needs of those with visual impairments [4], [5]. To address this, researchers have explored approaches that focus on generating image descriptions that emphasize the cognitive and visual details most relevant to individuals with visual impairments, rather than aiming to create comprehensive, story-like narratives [6]. These efforts have highlighted the importance of balancing the level of detail, the prioritization of key visual features, and the clear communication of salient information to ensure that the generated captions are maximally useful for the target audience [7].

Image caption generator now uses an optimized CNN-based encoder, RNN-based decoder model that replaces RNN and LSTM architectures with IndRNN, which learns longer-term dependencies more efficiently than LSTMs [8]. Using CNN-LSTM automatic picture captioning model a qualitative and accurate captions that explain the picture in natural language [9]. The functional API from keras allowed us to integrate our LSTM and image vector models to predict the next word in the sequence as our input was an image vector and a partial caption [10].

Vanilla-RNN predicts better by focusing on the relevant input word portion, according to research. Beam search outperforms greedy search for correlation evaluation [11]. The basics of certain photo captioning systems are covered. Provides field-specific data and assessment indexes. Previous photo captioning algorithms improved prediction but could not produce customised description statements [12].

We identified several picture captioning models and methods. CNN excels in picture content extraction, whereas RNN and LSTM are popular language creation models. LSTM outperforms RNN. Several research have employed encoder-decoder and attention methods for scene interpretation. MSCOCO is best for picture captioning since it comprises non-iconic images [13], [14]. Research provide visually impaired individuals with greater options to communicate and comprehend their surroundings. Allow visually challenged individuals to engage more intimately with others without worry of blurring or uncertainty [15]-[17].

Image captioning for both the original and affine altered images. Image hashing has been main method for this. Common picture captioning datasets and assessment measures are explained in this article. The suggested technique generates accurate and meaningful captions for various affine altered pictures. Model handled all transformations except rotation. To label all changed photos consistently, work accuracy must be improved [18]. Study introduces a semantic picture retrieval approach that uses produced textual descriptions to examine the high-level semantic content inside them. A comparison of actual and created descriptions for RS picture retrieval reveals a 0.3 mean bilingual evaluation understudy (BLEU) score difference. We want to enhance caption creation in future work to close the gap and enhance retrieval performance [19].

A network that generates numerous captions by matching the guide map to the image's area. The VAE encoder creates a latent space of CAM vectors. The latent space vector reflecting the image-attention region style is then retrieved. The VAE decoder uses this vector as CNN input and caption condition. While retaining accuracy, the suggested model outperforms the basic model in diversity [20].

The study presents an unsupervised approach for training an image captioning model without linked picture-sentence data, aiming to achieve three training objectives: captions should be indistinguishable from corpus sentences, the model should convey object information, and features should align in common latent space for bi-directional reconstructions [21].

The neural network automatically analyzes the image and provides a detailed description. The system uses convolutional neural networks to expand images and RNN to generate sentences. The model is trained to obtain the result of the proposed image. Tests of many materials demonstrate the consistency of quality and quantity measurements using BLEU (electronic quality measurement or machine translation). Experiments show that the proposed model performs better on large datasets for descriptive images. It would be interesting to investigate whether unsupervised information, including images and texts, can improve image identification strategies [22], [23].

This article explores deep-learning-based picture captioning algorithms, presenting a taxonomy of approaches, their benefits and drawbacks, evaluation metrics, and experimental outcomes. It discusses future study directions, highlighting the need for reliable, high-quality captions [24]. The transformer object relation transformer is a new image captioning model that investigates 2D location and size correlations among recognized items. It uses geographical connection information, improving spatial awareness. Future work will focus on object-word relationships, enhancing model performance and understanding ability [25]. The study employs a neural network to evaluate photos and create English captions. The model categorizes descriptions based on word proximity, with larger datasets improving model performance and reducing losses. The study also explores the potential of unsupervised data for improved caption creation [26].

Image captioning has improved with deep learning, enhancing accuracy and efficiency in various sectors like health, security, and military. This research also advances image annotation, VQA, cross media retrieval, and video captioning [27]. The research explores using VAE to create image captions using a variation of MNIST. The architecture, consisting of an RNN encoder and a fully linked network decoder, effectively generalizes to unseen captions. Similar tasks were found in literature, including arithmetic image production. Future research aims to test deconvolution layers and use attention mechanisms for iterative picture generation [28].

3. PROPOSED METHOD

The proposed method with basic system flow is shown in Figure 2. An image is used as input to our system, but it is not a raw image; instead, it undergoes a preprocessing step as shown in Figure 3. The processing method is as follows: we apply an autoencoder to the image. The autoencoder encodes and compresses the image, and the resulting compressed layer, also known as the bottleneck layer or latent space representation, is provided as input to the proposed model. Detail description with result of this method is described in in our prior publication [2].

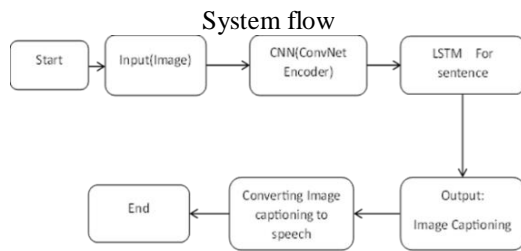


Figure 2. The proposed system's flow

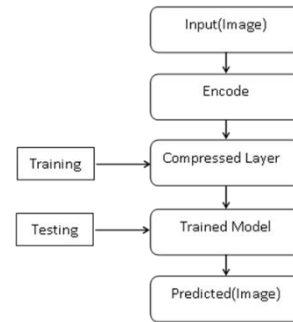


Figure 3. Applying autoencoder on image and uses a compressed layer for training

The proposed work describes the system that contains 4 modules with details of each as follows.

- The image to be processed is first provided to the image based module, which creates a vector known as the input image's feature vector using the CNN algorithm's convolutional and pooling layer. Each convolutional layer follows by a ReLu layer. The feature vector is then shrunk in size using the pooling layer before being sent to the following model. Since we just require the feature vector, the fully connected network, the final layer of CNN, is not included in our model. As feature extractors, convolutional and pooling layers are employed, whereas fully connected networks are used as classifiers.
- The next module, language based module, receives the output of the previous model, which is a vector of features generated. Using the LSTM algorithm-an advanced version of the RNN with the benefit of storing long sequences of data-the encoded features vector is decoded into a natural language caption. A memory cell in an LSTM allows data to be stored for an extended amount of time. To help the algorithm decide when to start and stop the sentence sequence, the sentence/caption sequence contains two unique tokens called startseq and endseq.
- Finally, the caption is produced. The caption model emphasizes relationships between items as well as objects, colors, and activities.
- Transforming the picture caption into audio.

Algorithm

Step 1: apply pre processing on input image.

Step 2: the input image's preprocessed pixel matrix is sent into an image-based model system, which creates features using CNN vector.

Step 3: features of output LSTM is used to decode vector into a natural language caption.

Step 4: convert this caption into speech.

Modules of system

- Step 1: apply pre processing on input image.

The images are not understood by the machine or system. First, the received image is converted into a fixed-sized pixel matrix ($224 \times 224 \times 3$), where the color code of each pixel is assigned to its proper spot. The noise in each and every picture is then eliminated during pre-processing. After the image has been converted to grayscale, a threshold value is determined to divide it into foreground and background. Every object in a picture undergoes edge detection. The output of the image pre-processing model and the input for the following module is the final pixel matrix.

- Step 2: the input image's preprocessed pixel matrix is sent into an image-based model system, which creates features using CNN vector.

A redesigned CNN module that uses the convolutional and pooling layers to extract features. The matrix of pixels that is the output of the previous pre-processing module serves as the input for the image-based module. This module extracts features from the image pixel matrix. The first CNN layer used in this module to extract features is the convolutional layer. Every convolutional layer is followed by a ReLU layer. The feature vector is then shrunk in size without sacrificing any of the image's features by applying a pooling layer. Features are retrieved from this module and stored in a feature vector. These features include objects, verbs that describe the object's behavior, the object's color, and the most significant relationship between the object. The feature vector, which is the output of this module, is the input of the next module. The vector's size is linearly converted to the LSTM network's input size, which is utilized in the next module.

- Step 3: features of output LSTM is used to decode vector into a natural language caption.

The linear feature vector for a particular input image serves as the input for the language-based module. This module's primary goal is to translate the encoded features into a plain language that users with LSTM can comprehend. The LSTM algorithm is used by the module because it resolves the variation gradient problem with the RNN algorithm and has the ability to hold a lengthy data sequence without losing it. The LSTM uses its memory cells to store the data. We must first pre-define our label and target text to train The LB, or language based model. Starting with a start token, the label saves the data in a series, and the Target stores the sequence with an end token.

Generation of captions the language based module comes first in the sequence of modules that make up the final module, caption generation. This module's objective is to use the guidelines from the previous module to write a caption for an input image in a linear fashion. This module ends with the generation of a caption in a format that is easily understood by humans.

- Step 4: convert this caption into speech

Lastly, captions based on text are converted into speech so visually impaired people can hear and understand easily.

The Figure 4 illustrates a detailed process flow for an image captioning system of proposed system that involves training, processing, and generating descriptions for images and finally converting to the speech. Here is a step-by-step explanation:

Training phase

- Train images: collect training images for the model.
- Resize image: resize the training images to a standard size of 299x299 pixels.
- Feature extraction model (Xception model): use the Xception model (a convolutional neural network) to extract features from the images. This model outputs dense feature representations.
- Train description: collect descriptions (captions) corresponding to the training images.
- Cleaning operations: perform cleaning operations on the training descriptions (e.g., removing punctuation, converting to lowercase).
- Generate vocabulary: generate a vocabulary from the cleaned descriptions.
- Create tokenizer for each word: create a tokenizer that converts each word in the descriptions to a numerical format.

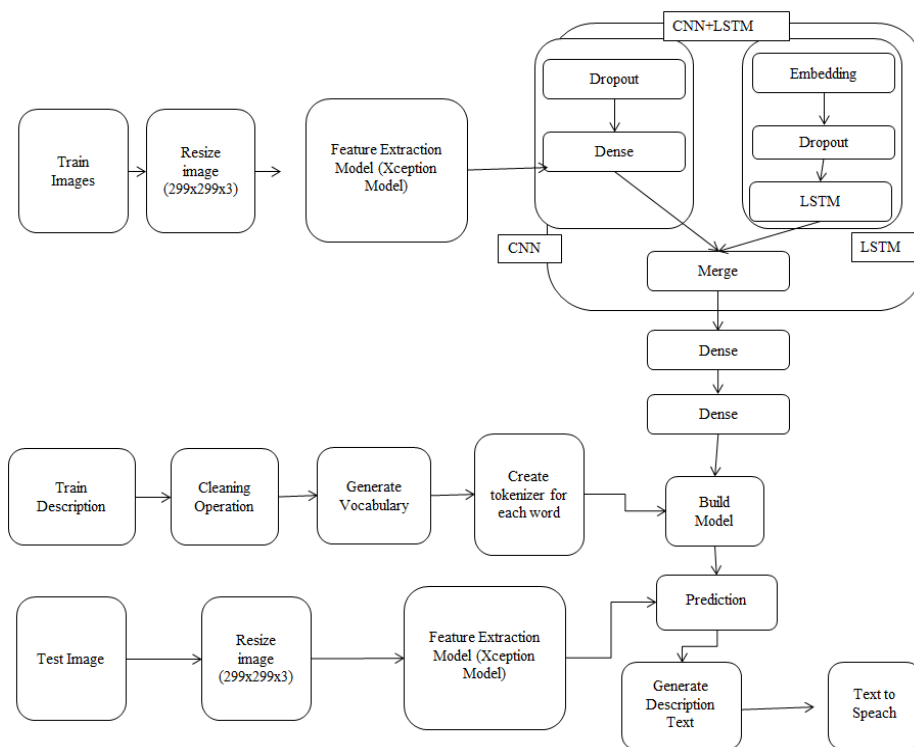


Figure 4. Specifications for the proposed system's flow

Model building phase: CNN + LSTM model (reprented in Figure 5)

- CNN: uses extracted image features from the Xception model.
- LSTM: processes the tokenized descriptions.
- Dropout: applies dropout regularization to prevent overfitting.
- Embedding: converts tokenized words into dense vector representations.
- Dense layers: fully connected layers to combine features from CNN and LSTM.
- Merge: combines outputs from CNN and LSTM.
- More dense layers: additional fully connected layers to further process the combined features.

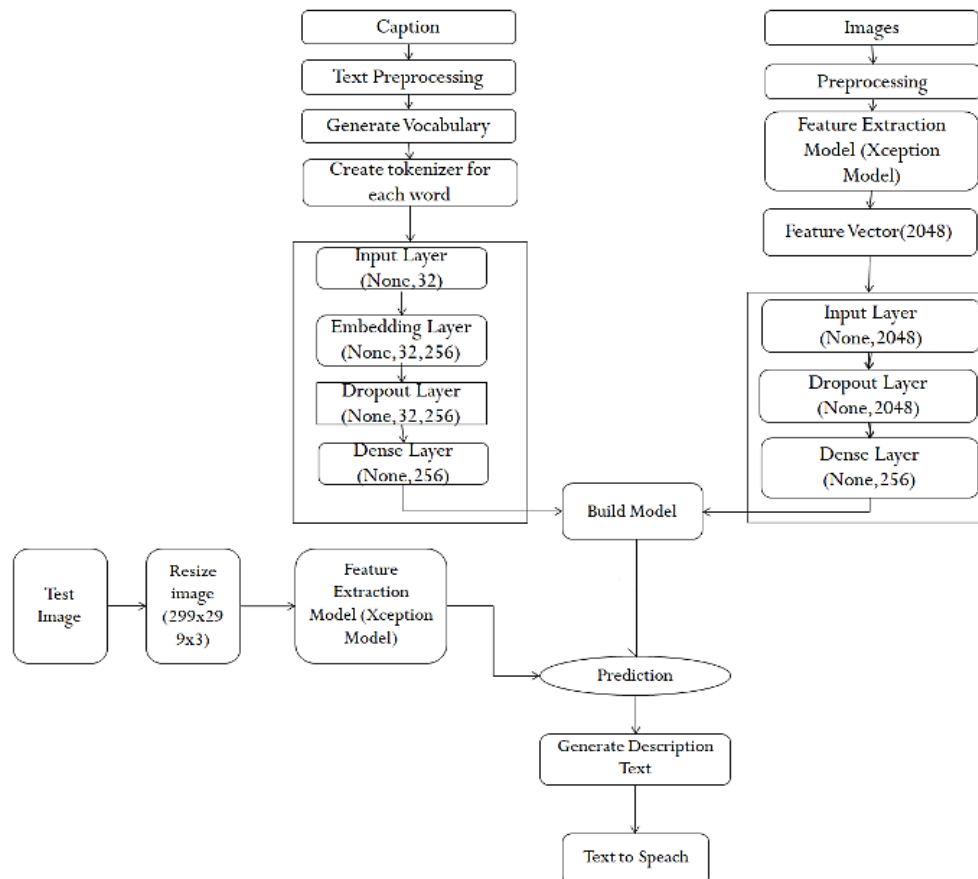


Figure 5. Detail understanding of CNN and LSTM model in proposed system

This Figure 5 outlines a detailed architecture for an image captioning system, highlighting both the training and prediction phases. Here's a step-by-step explanation:

Caption processing

- Caption: start with text descriptions of the images.
- Text preprocessing: clean and preprocess the text data.
- Generate vocabulary: create a vocabulary from the cleaned text data.
- Create tokenizer for each word: tokenize the text, converting words into numerical tokens.
- Input layer (None, 32): define the input layer for the text data, with a sequence length of 32.
- Embedding layer (None, 32, 256): embed the tokens into dense vectors with 256 dimensions.
- Dropout layer (None, 32, 256): apply dropout to the embedding layer to prevent overfitting.
- Dense layer (None, 256): use a dense layer to further process the embedded vectors.

Image processing

- Images: start with the image dataset.
- Preprocessing: preprocess the images to standardize them.
- Feature extraction model (Xception model): use the Xception model to extract features from the images.

- Feature vector (2048): extracted features are represented as vectors of size 2048.
- Input layer (None, 2048): define the input layer for the image features.
- Dropout layer (None, 2048): apply dropout to the input layer to prevent overfitting.
- Dense layer (None, 256): use a dense layer to process the feature vectors.

Model building

Build model: combine all the components, the processed text data and image features to build the final image captioning model.

Testing phase

- Test image: provide a test image for caption generation.
- Resize image: resize the test image to 299×299 pixels.
- Feature extraction model (Xception model): extract features from the test image using the Xception model.
- Prediction: use the built model in previous step to generate a description for the test image.
- Generate description text: convert the model's output into a descriptive text.
- Text to speech: optionally, convert the generated description text to speech.

This image captioning system involves preprocessing both text descriptions and images, extracting features from images using the Xception model, and processing text using embedding and dense layers. The combined model generates text descriptions for new images, which can be converted to speech. The flow includes handling both training data (captions and images) and test data (new images for generating descriptions).

4. EXPERIMENTAL RESULTS

The suggested approach can be validated using the BLEU metric as shown in Table 1, which is widely used to evaluate the quality of machine-generated translations or text against reference descriptions. The core idea behind the BLEU metric is the calculation of precision, which measures how well the words or phrases (n-grams) in the candidate sentence match with those in the reference sentence.

The precision is calculated by dividing the number of matching n-grams (continuous sequences of words) between the generated sentence (G) and the reference sentence (R) by the total number of n-grams in the candidate sentence. In other words, it checks how many n-grams in the generated output also appear in the reference text. The BLEU score is computed by taking the geometric mean of the n-gram precision values across different n-gram lengths (e.g., unigrams, bigrams, trigrams). The geometric mean, denoted as:

$$P(N, G, R) = (\prod_{n=1}^N P_n)^{1/N}$$

here, P_n represents the precision for n-grams of length n, and N is the maximum n-gram length (e.g., N=4 typically represents 1-gram to 4-gram precision).

$$P_n \text{ is calculated as: } P_n = m_n / l_n$$

Where: m_n is the number of matching n-grams between G and R, l_n is the total number of n-grams in G. However, to avoid favoring overly short generated sentences, a brevity penalty (BP) is introduced. The brevity penalty ensures that if the generated sentence is too short compared to the reference sentence, the BLEU score is penalized. The BP is defined as:









$$BP(G, R) = \min\left(1.0, \exp\left(1 - \left(\frac{\text{len}(R)}{\text{len}(G)}\right)\right)\right)$$

where: $\text{len}(R)$ is the length of the reference sentence, $\text{len}(G)$ is the length of the generated sentence. If the generated sentence is longer than or equal to the reference sentence, BP is set to 1 (i.e., no penalty). Otherwise, it applies a penalty that decreases the BLEU score when the candidate sentence is too short. Therefore, the final BLEU score is calculated as below.

$$BLEU = P(N, G, R) \times BP(G, R)$$

An important note is that when higher-order n-gram precision (e.g., 4-gram precision for $n=4$) is zero, the entire BLEU score will be zero, even if lower-order n-grams (e.g., unigrams, bigrams) match. This is because BLEU emphasizes not only individual word matches but also the coherence of longer sequences. As a result, BLEU captures both exact matches and fluency in the generated text.

Table 1. Experimental results for CNN+LSTM on flicker dataset

Image	Prediction	BLEU	BLEU1	BLEU2	BLEU3	BLEU4
	Dog is running through the snow	56.2341	50	20	100	100
2319175397_3 e586cfaf8.jpg						
	Two children are playing with sparklers in the grass	40.4289	29.82	11.18	89.48	89.48
3477712686_8 428614c75.jpg						
	Man in black shirt and black pants and beard is standing in front of crowd	34.378	29.17	6.25	87.16	87.16
3477712686_8 428614c75.jpg						
	Man in red shirt is walking on top of snow	38.6097	80	22.22	12.5	100
3139393607_f 0a54ca46d.jpg						
	Boy in blue bathing suit jumps into pool	51.6973	50	14.28	100	100
3621652774_f d9634bd5b.jpg						
	Man in red shirt is jumping on the edge of the ocean	75.9836	33.33	100	100	100
2505056124_1 276e8dbcb.jpg						
	Two small dogs are playing with ball	61.4788	14.29	100	100	100
2229179070_d c8ea8582e.jpg						
	Dog is running through the grass	60.4275	66.67	20	100	100
123889082_d3 751e0350.jpg						
	Man in red shirt is walking on top of large rock	82.1097	45.45	100	100	100
2752329719_8 68545b7d2.jpg						

5. CONCLUSION




In conclusion, the proposed image captioning system effectively amalgamates diverse sophisticated machine learning models to produce informative captions for pictures. The system utilizes a CNN-based image model to process input pictures, extract features, and subsequently use an LSTM-based language model to convert those characteristics into natural language captions. The method proficiently tackles both feature extraction and sequence creation issues, guaranteeing precision and fluency. The system also has a text-to-speech feature to render the generated captions accessible for visually challenged users. The BLEU measure is utilized to assess the system's performance, confirming that the generated captions closely align with human-written descriptions. This method employs multi-layered modules to deliver an efficient and scalable solution for practical applications in picture captioning.

Our model, which is based on multi-label classification using CNN and quick text to speech, is helpful for extracting objects from images and creating captions depending on the datasets that are supplied. Here for creating image captions, method used includes concept of RNN, LSTM, and CNN. As input is not a raw image but a processed image which actually helps for better and fast results. Further as a future work summarization of live video of surrounding of visually impaired people can be done through which a continuous description of the spatial awareness, distances. Hence enabling visually impaired individuals to navigate their surroundings independently and confidently.




REFERENCES

- [1] M. Kouthair Khribi, "Toward accessible online learning for visually impaired and blind students," *Nafath*, vol. 7, no. 19, Jan. 2022, doi: 10.54455/mcn.19.02.
- [2] N. B. Shah and A. P. Ganatra, "Evaluation of autoencoders: training using original, encoded and decoded images for prediction," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 12, no. 3, pp. 2563–2569, 2024.
- [3] N. Shah and A. Ganatra, "comparative study of autoencoders-its types and application," *6th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2022 - Proceedings*, pp. 175–180, 2022, doi: 10.1109/ICECA55336.2022.10009387.
- [4] H. MacLeod, C. L. Bennett, M. R. Morris, and E. Cutrell, "Understanding blind people's experiences with computer-generated captions of social media images," in *Conference on Human Factors in Computing Systems - Proceedings*, May 2017, vol. 2017-May, pp. 5988–5999, doi: 10.1145/3025453.3025814.
- [5] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, 2018, vol. 1, pp. 2556–2565, doi: 10.18653/v1/p18-1238.
- [6] D. L. Fernandes, M. H. F. Ribeiro, F. R. Cerqueira, and M. M. Silva, "describing image focused in cognitive and visual details for visually impaired people: an approach to generating inclusive paragraphs," in *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2022, vol. 5, pp. 526–534, doi: 10.5220/0010845700003124.
- [7] V. Wadhwa, B. Gupta, and S. Gupta, "AI based automated image caption tool implementation for visually impaired," in *ICIERA 2021 - 1st International Conference on Industrial Electronics Research and Applications, Proceedings*, Dec. 2021, pp. 1–6. doi: 10.1109/ICIERA53202.2021.9726759.
- [8] S. Rampal, S. Gupta, S. Verma, and D. K. Vishwakarma, "Image captioning using IndrRNN," *International Journal of Advanced Science and Technology*, vol. 29, no. 08, pp. 2211–2217, 2020.
- [9] P. Raut and R. A. Deshmukh, "An advanced image captioning using combination of CNN and LSTM," *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, vol. 12, no. 1S, pp. 129–136, Apr. 2021, doi: 10.17762/turcomat.v12i1s.1593.
- [10] M. Raypurkar, A. Supe, P. Bhumkar, P. Borse, and S. Sayyad, "Deep learning based image caption generator," *International Research Journal of Engineering and Technology*, vol. 08, no. 03, pp. 554–559, 2021.
- [11] S. H. Choi, S. Y. Jo, and S. H. Jung, "Component based comparative analysis of each module in image captioning," *ICT Express*, vol. 7, no. 1, pp. 121–125, Mar. 2021, doi: 10.1016/j.ict.2020.08.004.
- [12] C. Wang, Z. Zhou, and L. Xu, "An integrative review of image captioning research," *Journal of Physics: Conference Series*, vol. 1748, no. 4, p. 042060, Jan. 2021, doi: 10.1088/1742-6596/1748/4/042060.
- [13] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor, and M. Khan, "Image captioning using deep learning: a systematic literature review," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 278–286, 2020, doi: 10.14569/IJACSA.2020.0110537.
- [14] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: transforming objects into words," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [15] H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation methods," *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–13, Jan. 2020, doi: 10.1155/2020/3062706.
- [16] P. Waghmare and D. S. Shinde, "Artificial intelligence based on image caption generation," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3648847.
- [17] J. Y. Lin, C. L. Chiang, M. J. Wu, C. C. Yao, and M. C. Chen, "Smart glasses application system for visually impaired people based on deep learning," in *Indo - Taiwan 2nd International Conference on Computing, Analytics and Networks, Indo-Taiwan ICAN 2020 - Proceedings*, Feb. 2020, pp. 202–206. doi: 10.1109/Indo-TaiwanICAN48429.2020.9181366.
- [18] M. Nivedita and Y. A. V. Phamila, "Image captioning for affine transformed images using image hashing," *International Journal of Engineering and Advanced Technology*, vol. 9, no. 1, pp. 4736–4741, Oct. 2019, doi: 10.35940/ijeat.A2022.109119.
- [19] G. Hoxha, F. Melgani, and B. Demir, "Retrieving images with generated textual descriptions," in *International Geoscience and Remote Sensing Symposium (IGARSS)*, Jul. 2019, pp. 5812–5815. doi: 10.1109/IGARSS.2019.8899321.
- [20] B. Kim, S. Shin, and H. Jung, "Variational autoencoder-based multiple image captioning using a caption attention map," *Applied Sciences (Switzerland)*, vol. 9, no. 13, p. 2699, Jul. 2019, doi: 10.3390/app9132699.
- [21] Y. Feng, L. Ma, W. Liu, and J. Luo, "Unsupervised image captioning," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Jun. 2019, vol. 2019-June, pp. 4120–4129, doi: 10.1109/CVPR.2019.00425.
- [22] B. Haripriya, S. Haseeb, G. M. Srushti, and M. Prakash, "Image captioning using deep learning," *International Journal of Engineering Research & Technology*, vol. 8, no. 5, pp. 381–388, 2019, doi: 10.17577/IJERTV8IS050284.
- [23] L. Srinivasan, D. Sreekanthan, and A. A. L., "Image captioning - a deep learning approach," *International Journal of Applied Engineering Research*, vol. 13, no. 9, pp. 7239–7242, 2018.
- [24] R. Staniute and D. Šešok, "A systematic literature review on image captioning," *Applied Sciences (Switzerland)*, vol. 9, no. 10, p. 2024, May 2019, doi: 10.3390/app9102024.
- [25] M. D. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, Nov. 2019, doi: 10.1145/3295748.
- [26] C. Amritkar and V. Jabade, "Image caption generation using deep learning technique," in *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018*, Aug. 2018, pp. 1–4, doi: 10.1109/ICCUBEA.2018.8697360.
- [27] S. Liu, L. Bai, Y. Hu, and H. Wang, "Image captioning based on deep neural networks," *MATEC Web of Conferences*, vol. 232, p. 01052, Nov. 2018, doi: 10.1051/mateconf/201823201052.
- [28] N. A. Gontier, J. Romoff, and P. Parthasarathi, "Variational encoder decoder for image generation conditioned on captions," *33rd International Conference on Machine Learning*, vol. 48, 2016.

BIOGRAPHIES OF AUTHORS

Nidhi B. Shah    working as an assistant professor at computer engineering Department Sardar Vallabhbhai Patel Institute of Technology-Vasad (SVIT) Gujarat. She is part time research Scholar in department of computer engineering, Charusat University Gujarat. She has received M.Tech. degree in computer engineering from Dharamsinh Desai University (DDU), Nadiad Gujarat in 2011. She has received B.E. Degree in Computer Engineering from Sardar Vallabhbhai Patel Institute of Technology-Vasad, Gujarat 2007. Her area of interest are data mining, machine learning, and deep learning. She can be contacted at email: nbshah999@yahoo.com.



Amit P. Ganatra    is Provost. At Parul University Vadodara India. His citations are 2589, h-index-23, i-10 index- 43, SCOPUS-80 documents He has guided 100+ Dissertations, guided more than 100+ industry projects at undergraduate level, he is guided 12+ Ph.D. students. He is an active member of IEEE (senior member), ACM and CSI professional society chapters. His is a reviewer in many Journals and Conferences. His area of interest includes database technologies, data mining, business intelligence and data analytics, artificial intelligence, machine learning, cloud computing, IoT, block-chain technologies, system software, and software engineering. He can be contacted at email: provost@paruluniversity.ac.in.