International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Deep Neural Networks for Automated Image Captioning to Improve Accessibility for Visually Impaired Users

[1]Dr. Yashwant Dongare, [2]Dr. Bhalchandra M. Hardas, [3]Dr. Rashmita Srinivasan, [4]Dr. Vidula Meshram, [5]Dr. Mithun G. Aush, [6]Dr. Atul Kulkarni

Submitted: 17/07/2023 **Revised**: 08/09/2023 **Accepted**: 27/09/2023

Abstract- Many researchers are using artificial intelligence and machine learning models to aid the blind due to the advancements in image understanding and automatic image captioning. This research investigates the design and evaluation of deep neural network models for automatic picture captioning, with a focus on improving accessibility for those with visual impairments. The recommended method makes use of deep learning techniques, specifically convolutional neural networks (CNNs) for identifying characteristics in images and recurrent neural networks (RNNs) for generating descriptive captions. The appropriate features are extracted from the input photographs by the CNN and supplied into the RNN so that textual descriptions can be generated. The models are created utilizing techniques like attention processing and beam search to improve the caliber and coherence of the output captions. They are trained using large-scale image caption datasets. Extensive tests are carried out utilizing benchmark datasets as MS COCO and Flickr30k to assess the performance of the created models. The effectiveness of the generated captions is evaluated using objective measures like BLEU, METEOR, and CIDEr. Additionally, a user research with people who are visually impaired is carried out to determine how well the automatic picture captioning system improves accessibility. The outcomes show that the suggested deep neural network models for automatic picture captioning are effective.

Keywords*: Image caption, Convolution neural network, deep learning, LSTM, RNN, Automated caption generation*

I. Introduction

Images and videos play a crucial role in informational communication and in enhancing our comprehension of the world around us. Accessing and understanding visual information, however, presents major difficulties for people with vision impairments. Alternative text descriptions, a common technique for enabling accessibility, frequently fall short of fully encapsulating the context and specifics of visual content. Deep neural network-based automated image captioning systems have been developed in response to this issue, and they present a promising way to improve accessibility for users who are blind or visually impaired.

Creating textual descriptions that accurately reflect the subject matter and setting of an image is known as automated image captioning. These systems may automatically assess visual data and provide captions that provide a thorough comprehension of the image. Automated image captioning has several potential uses for enhancing accessibility. To navigate online environments and carry out daily operations like online shopping or using banking

[1]*Assistant Professor, Computer Engineering, Department Vishwakarma Institute of Information Technology Pune, Maharashtra, India*
yashwant.dongare@viit.ac.in
[2]*Assistant Professor, Department of Electronics and Computer science, Shri Ramdeobaba college of Engineering and Management, Nagpur, Maharashtra, India*
hardasbm@rknec.edu
[3]*Associate Professor, Department of Civil Engineering, Maharashtra Institute of Technology (Autonomous), Aurangabad, Maharashtra, India*
srinivasan.rashmita@gmail.com
[4]*Assistant Professor, Department of Computer Engineering, Vishwakarma Institute of Information Technology Pune, Maharashtra, India*
vidula.meshram@viit.ac.in
[5]*Assistant Professor, Department of Electrical Engineering, Chh. Shahu College of Engineering, Aurangabad, Maharashtra, India*
mithun.csmss@gmail.com
[6]*Professor, Department of Mechanical Engineering, Vishwakarma Institute of Information Technology Pune*
atul.kulkarni@viit.ac.in

services, blind people primarily rely on web accessibility standards. Although it has long been standard practice, alternate text (alt text) descriptions of photographs frequently fall short of giving a comprehensive understanding of the visual content [5]. Although identifying images on the web presents considerable hurdles, blind people's use of image captions to access the internet is an essential aspect of their everyday life [4]. For those who are blind, carrying out necessary tasks like internet banking or food shopping becomes difficult [15]. Given that the web is a crucial resource for the blind, it is crucial to follow best practices for web accessibility by include alternate text (alt text) descriptions for images.

Deep neural network models for automated picture captioning have been created and thoroughly researched to overcome this constraint. By skilfully fusing language modelling capabilities of RNNs with the visual information gathered from images using CNNs, these models seek to provide precise and contextually relevant captions. This research presents a thorough analysis of the design and assessment of deep neural network models for automatic picture captioning, with a focus on their potential to increase accessibility for users who are visually impaired. We investigate cutting-edge methods to improve the quality and coherence of generated captions, including attention processes and beam search. Additionally, we thoroughly evaluate the created models' performance using benchmark datasets and objective measures. The results of this study help close the accessibility gap for people who are blind or visually impaired by utilizing deep neural networks for automatic picture captioning. These systems have the potential to enable visually impaired individuals to access and grasp a wide range of visual information by offering more thorough and contextually appropriate explanations of visual content.

The photos' captions enable blind people to participate in social activities, access more information online, and aid in product purchases. Blind persons can learn more about the photographs because to the automatically generated captions [18]. Researchers have recently concentrated on improving web accessibility using various methods. These techniques can be divided into three groups: machine-based, hybridized technologies, and crowdsourcing.

II. Related Work

Song et al.'s [29] research created the ground breaking Visual Text Merging (VTM) framework for creating image descriptions. Their strategy entailed creating an attention model to precisely interpret visual information and produce illustrative captions. They suggested an adaptable VTM network dubbed avtmNet to combine the visual and textual information. The performance of the merging network was tested in tests utilizing well-known datasets as COCO2014 and Flickr30K. The outcomes showed how well the suggested method worked to precisely merge visual and text data, resulting in excellent image captions.

A thorough assessment on deep learning approaches and their difficulties in discovering items to market to people who are visually impaired (VI) was carried out by Wei et al. [30]. Blind people encounter many difficulties in their daily lives, especially when doing things like shopping. The researchers explored deep learning frameworks that can help in reliably identifying products through image captions in order to assist VI people. They used a number of datasets, including “GroZi-120, RPC, D2S, and Groci-3.2 K”. For people with disabilities (VI), the correctness of captions is crucial since errors can seriously impair their capacity to make wise purchasing decisions. The study examined multiple deep learning frameworks that accurately identify products and produce captions for images.

Ensemble Caption (EnsCaption), a ground-breaking framework that integrates caption production and caption retrieval algorithms for image captioning, was introduced by Yang et al. in their work [34]. By utilizing both caption creation and retrieval procedures, this strategy strives to improve the accuracy of caption generation. Using the Flickr-30K dataset, the researchers assessed their evaluation approach. The findings showed that the newly developed method increases the precision of caption production, producing more accurate and contextually appropriate captions for photos. By combining the capabilities of caption production and retrieval, Yang et al.'s Ensemble Caption framework offers a revolutionary method for captioning images, ultimately enhancing the overall quality and accuracy of generated captions.

Their [36] framework made use of the provided classifier to produce captions that matched the subjects of the pictures. The image and its connected topic served as the input for this devised method, while the generated image caption served as the output. To achieve efficient information representation, the framework used an embedding technique and a hierarchical structure. A bidirectional caption image retrieval method was used to find picture captions. Their research's findings showed that the created method produced captions for images of a higher caliber. Using a hierarchical architecture and a CNN-based multiple labeling classifier, Yu et al. enhanced performance in producing picture captions that precisely matched the images.

A trainable end-to-end framework for creating picture captions was proposed by Iwamura et al. [10] in their study. They used three datasets: copyright-free photos, MSCOCO, and MSRVTT2016-image. "*Feature extraction, motion estimation, object detection, and caption synthesis*" were the four distinct stages of the framework. The researchers created a motion-CNN model that automatically retrieved motion-related data from the input photos in order to extract motion features. The photos were processed by the CNN module to extract visual features, which were then sent to the object detection stage for object recognition.

The attention features for the caption generating component. These attentional characteristics were fed into an LSTM network together with the retrieved visual and motion features. The LSTM network produced descriptive captions for the input images by using the correlated attention features. Iwamura et al. [10] sought to automate the entire process of picture captioning, from feature extraction to caption production, by merging these four steps into an end-to-end trainable framework. This method made it possible to represent the input images completely and contextually, which resulted in image descriptions that were more precise and detailed.

Table 1: Summary of related work in image caption generation

Paper	Used Dataset	Method used	Remark
[1]	COCO2014 and Flickr30K	avtmNet	Image Resizing
[3]	COCO2014 and Flickr30K	DenseNet+LSTM	Image Resizing
[2]	"RPC, D2S, Grozi-120, Groci-3.2 K"	Deep learning models	"Noise elimination and removing redundant data"
[4]	MS COCO and Flickr30K	Dual LSTM	-
[5]	Dataset for Hindi Genome	LSTM Model and CNN Model	IndicNLP Tokenizer
[6]	"MSR-VTT2016-Image, MSCOCO"	CNN_LSTM	"Image Resizing"

III. Dataset Description

Performance of the suggested technique is assessed using Freiburg Groceries and Grocery Store Datasets, two publically accessible datasets. Examples of photographs from the "Freiburg Groceries Dataset" are shown in Figure 1. MS COCO dataset with Freiburg Groceries, the initial dataset, has a total of 4947 photos divided into 25 categories. These photos were gathered in Germany utilizing phone cameras from residences, workplaces, and retail locations. A total of 4947 pictures were processed for this investigation, with 3462 utilized for training and 1485 used for testing.

Fig 1: Dataset sample images (MS COCO Dataset)

A popular benchmark for creating image descriptions from sentences is the Flickr30k dataset. An expanded version of the Flickr30k dataset called Flickr30k Entities is presented in this study. This enhanced dataset consists of 158k Flickr30k captions that have been enhanced with 244k reference chains. These links are made between instances of the same entities mentioned in several captions for the same image. The collection also includes 276k bounding boxes with handwritten annotations connected to these reference links. These annotations are essential for developing grounded language understanding and automatic image description. They aid in the creation of a new benchmark that is aimed primarily at localizing textual entity references within images.

Fig 2: Sample images of Dataset Flicker30

IV. Proposed Methodology

Simpler image captions are provided for photos without captions via automatic image captioning (AIC). The “data collection, non-captioned image selection, extraction of appearance and texture attributes”, and development of “automatic image captions” are integrated to process the suggested AIC using the EACNN model. The ARO method is used to choose uncaptioned images from the database in the first step. In order to extract look and texture information from uncaptioned photographs, the “spatial derivative & multi scale” (SDM) feature and “weighted patch based local binary pattern” (WPLBP) are utilized. Additionally, the extracted attributes are applied to distinguish the photographs with accuracy.

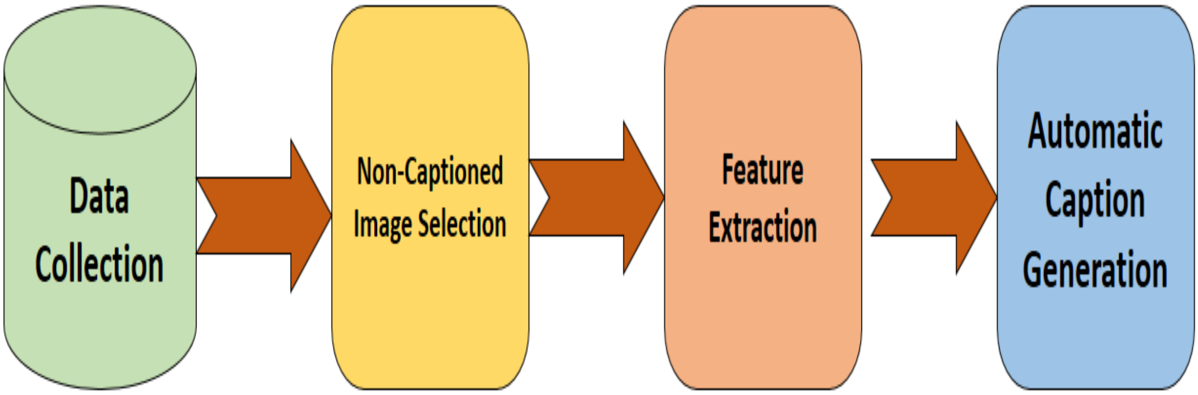


Fig 3: Block diagram of proposed model

A. Data Collection:

Data collection is the process of acquiring information from any publicly accessible internet source that allows us to assess the findings. A methodical approach to gathering information from several sources in order to produce an accurate and comprehensive result for a given topic of interest. For the research to remain authentic and for quality assurance to be ensured, precise data gathering is crucial. The “Freiburg Groceries Dataset and Grocery Store Dataset”, which are used for analysis and implementation in this work, are two separate datasets that were obtained from separate web sources. These online grocery datasets contain several classes that encompass a variety of goods and object kinds.

B. Image Selection with non-Captioned:

There are captioned and uncaptioned photographs in the dataset's raw image input. The primary goal of this effort is to automatically create captions for the photographs without captions. Here, only the uncaptioned photographs are first chosen using the “adaptive rain optimization” (ARO) algorithm. The selection of non-captioned photographs using fuzzy c-means addresses the issue of learning from an unbalanced dataset and is widely applicable in many academic areas, including engineering and image processing. Clustering, which divides the photos into two categories, such as captioned and uncaptioned photographs, is the fundamental concept underpinning Fuzzy C Means (FCM). Each piece of data can be connected to more clusters thanks to the unsupervised clustering approach used by the FCM algorithm.

Let us say that the input data points are given as P = p1, p2,..., pn, which is then broken down into CK clusters as CK = c1, c2,..., ck, where n denotes the number of components and CK denotes the number of pre-defined clusters. The objective function can be minimized as follows:

$$J\,min = \frac{1}{4}\sum C\ \text{i¼}\ 1\ \sum n\ \text{j¼1}\ Ug\ ijD2 \qquad (1)$$

where the membership degree is Uij, g is the fuzzification element, and Dij is the Euclidean distance from the jth instance to the ith cluster centroid. Equation (1) is minimized in light of the following circumstances:

$$1 \leq\ j \leq n :\ \sum C\ \text{i¼1}\ Uij\ \text{¼}\ 1 \qquad (2)$$

$$1 \leq i \leq C :\ \sum n\ \text{j¼1}\ Uij\ >\ 0 \qquad (3)$$

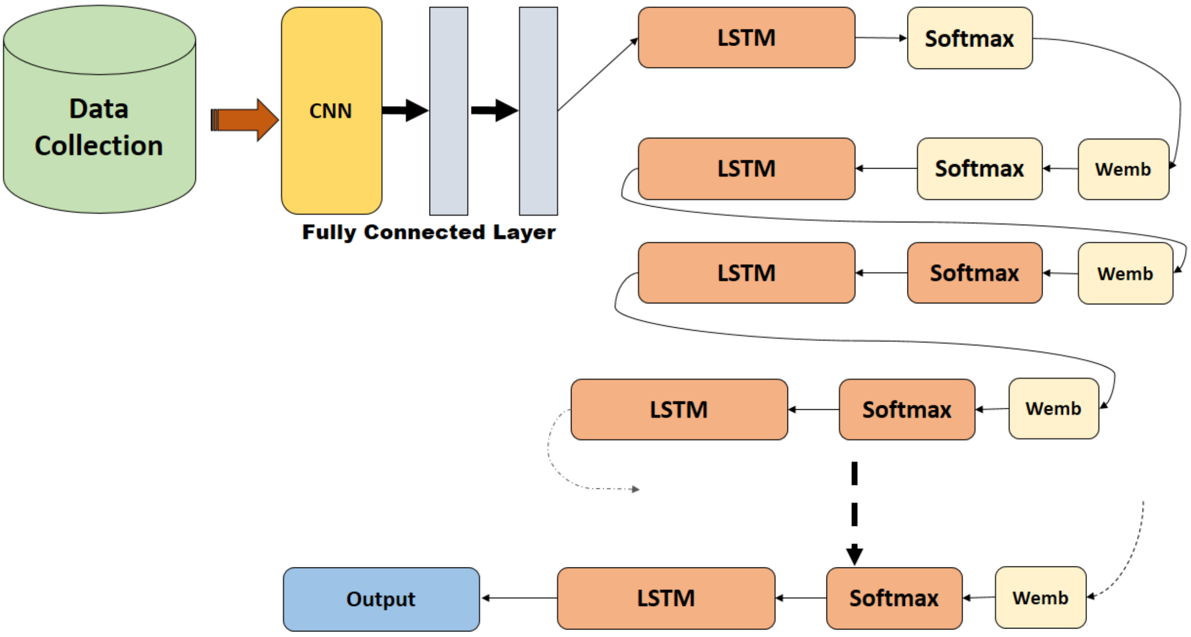


Fig 4: Representation of proposed CNN Model

Here, the derivative for both Uij and Cj parameters is set to zero in order to minimize Jmin using the Lagrange function.

$$Jmin = \sum C\ \text{i¼1}\ \sum n\ \text{j¼1}\ Ug\ ijD2\ ij\ \text{þ}\ \sum n\ \text{j¼1}\ \lambda k\ \sum n\ \text{i¼1}\ Uij - 1 \qquad (4)$$

The value of the fuzzification element is g = 2. The fuzzy set is used to generalize the FCM model, and the initial results are generated using fuzzy clustering with the appropriate competence and simplicity.

The Lagrange function k in the FCM model is optimized with the “RO algorithm”, which resolves the problems with “longer computation times, sensitivity to noisy data, and initialization”. The “RO algorithm” mimics the actions of raindrops. The initial population is formed using each solution as a raindrop. The primary function of RO is radius, and the population is initialized. The raindrops' R1, R2, and R connections are characterized thus,

$$R = (R_1^m - R_2^m)^{\frac{1}{4}} \quad (5)$$

where m is the factors that affect each droplet us count. In order to build the cluster group (“captioned and non-captioned image group”) quickly and with noise reduction, the number of total iterations is increased.

C. Feature Extraction:

The process of extracting visual information from a picture for indexing and retrieval is known as feature extraction. In order to extract pertinent information from the image during this procedure, texture and appearance elements are quite important. In order to improve human comprehension of the image, feature extraction aims to produce a set of features that are informative, non-redundant, and permit further learning.

To extract appearance-based features, a differential feature is a vector connected to a particular point in an image. The image's spatial derivatives are used to compute this feature. Lower order derivatives can be used as a feature for an image I with a point w, and they can be written as follows:

$$H = [\frac{\partial I}{\partial x}(w), \frac{\partial I}{\partial y}(w), \frac{\partial 2I}{\partial x2}(w), \frac{\partial 2I}{\partial xy}(w), \frac{\partial 2I}{\partial y(w)}] \quad (6)$$

Derivatives can be used to extract important statistical data from an image. Second-order derivatives are used to show bars or other structural components within the image, whilst first-order derivatives give insight into the intensity edge ness or gradient of the image.

The Local Binary Pattern (LBP) technique compares pixel changes within smaller parts of a picture to extract local information. The Directional Gradient (DG) information is then represented using '0' or '1' bits in a shorter binary string using the local information. Every center pixel in a circle's neighbourhood (referred to as zcn) is taken into account while extracting texture features using LBP, and it depends on the nearby pixels (referred to as zp, where p = 0, 1,..., P 1) with a given radius (R). This can be said in the following way:

$$LBP = \sum (zp - zcn) * 2^\wedge p \quad (7)$$

where the sum is taken from p = 0 to P-1.

WPLBP (“Weighted Patch Local Binary Pattern”), an upgraded or expanded version of LBP, makes use of a pyramidal structure. In WPLBP, a kernel sp from the set Sk is incorporated, and this kernel is added to the zpk patches from Zk. WPLBP's expression can be written as follows:

$$WPLBP = \sum_{k=1}^{k} f \left(\sum (zp, k-1 * sp, k)^2\right) \quad (8)$$

The N-dimensional “Gradient Descriptor” (GD) is used to build the training sets. In the WPLBP approach, a collection of patches Zt with t = 1, 2,..., T are used to calculate gradients using the Sobel operator. From the scaled image, these patches can be randomly selected. The image is scaled using the WPLBP method to gather patches for a training set, and then the texture feature is retrieved.

D. Automated Caption Generation:

Automatically creating captions for photos is known as image captioning. Given the enormous amount of picture data produced every day, a single image can hold a substantial amount of data. The next stage is to automatically create captions for photos without captions when relevant information has been extracted from an image. Convolutional Neural Networks (CNNs) are used in this procedure to automatically create image captions. The CNN model makes use of deep learning by combining the CNN and LSTM architectures. Reusing previously written captions for the target image using reverse image search lowers captioning errors. The AAS algorithm is used to increase accuracy even further. This deep learning-based caption production approach uses CNN to automatically annotate photos by reusing previously created captions. This reduces human error while still producing captions that are appropriate for visually impaired people.

The hybrid CNN/LSTM combination shown in Figure 4 represents the suggested architecture. Two popular deep learning architectures are CNN and LSTM. In this architecture, CNN models are used to create captions, while LSTM models use the reverse image search technique to choose the caption that is most likely to describe the target image. The LSTM network makes use of the characteristics of the training set to capture both short-term and long-term dependencies. The proposed model makes use of the capabilities of all layers to learn the internal time-series representation of the data. A typical Artificial Neural Network (ANN) model for image processing is CNN, usually referred to as ConvNet. To create pre-trained picture captions, CNN models accept image inputs and apply biases and learnable weights.

V. Proposed Algorithm

A) Convolution Neural Network:

Convolutional Neural Networks (CNNs) are a category of deep learning models that are frequently employed in the realm of image processing, encompassing tasks such as image classification, object detection, and image segmentation. Convolutional neural networks (CNNs) are highly suitable for these specific tasks due to their inherent ability to autonomously acquire hierarchical features from unprocessed pixel data.

Convolutional Neural Networks (CNNs) generally comprise a series of layers, which commonly include convolutional layers, pooling layers, and fully connected layers. Convolutional layers are utilized to perform the convolution operation, which aids in the acquisition of diverse image features. On the other hand, pooling layers are employed to downsample the feature maps, thereby mitigating the computational complexity involved. The final layers of a neural network, known as fully connected layers, are responsible for carrying out classification or regression tasks. Eq.1 represent the CNN operation

$$(f * g)(x, y) = \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} f(i, j).\, g(x - i, y - j)$$

where, f= “input image”, g= “filter”, (x,y)= “coordinates of the output feature map”, i and j= “variable denoted the coordinates”, $x - i, y - j$ = “spatial offsets used to align the kernel with the input feature map”

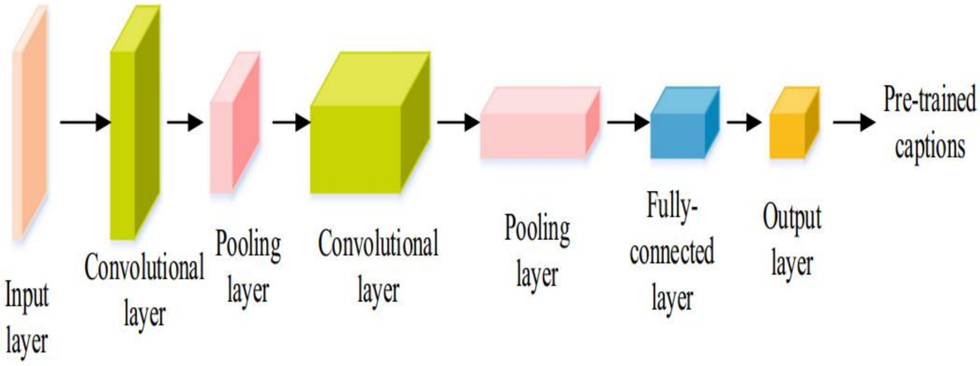


Fig 5: Structure of CNN

B) LSTM Algorithm:

The Long Short-Term Memory (LSTM) is a specific architecture of a recurrent neural network (RNN) that has been developed to effectively capture and represent sequential data that exhibits long-range dependencies [18]. LSTM models have demonstrated notable efficacy in tasks that involve the analysis of time series data, natural language processing, and speech recognition. This effectiveness can be attributed to their inherent capacity to effectively process and retain information across prolonged temporal intervals. The following is a concise elucidation eq. of LSTM

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$$

Where, f_t = “forget gate at time step t”, σ = “sigmoid activation function”, W_f and b_f= “weight and bias associated with forget gate”, h_{t-1}, x_t = “concatenation of previous cell state and current input”

The fundamental concept underlying LSTM is the incorporation of gated units, which enable the network to selectively acquire, retain, and discard information from previous time steps. This characteristic renders LSTM particularly suitable for tasks that heavily rely on context and memory. The forget gate is a crucial element within the LSTM cell, as it plays a pivotal role in determining which information from the preceding time step should be disregarded or preserved.

The forget gate generates values ranging from 0 to 1 for every element of the cell state. A value of 0 indicates complete forgetting, while a value of 1 signifies complete remembrance. The aforementioned values are employed to modify the cell state through the process of element-wise multiplication with the existing cell state.

LSTM models are equipped with additional gates, namely the input gate and output gate, that regulate the ingress and egress of information to and from the cell state. The inclusion of gates, in conjunction with the cell state, empowers LSTM networks to effectively capture extended dependencies in sequential data. This is achieved by granting the network the ability to make informed decisions regarding the retention and propagation of valuable information over time, while simultaneously discarding irrelevant information.

C) Proposed Hybrid CNN + LSTM Algorithm:

- The integration of Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks is frequently employed in scenarios that require simultaneous processing of spatial and sequential data. The present architectural design incorporates CNN for the purpose of capturing spatial characteristics from the input frames. LSTM networks are utilized to effectively model the temporal dependencies that exist across these frames. The hybrid architecture, which combines CNN and LSTM, comprises two primary components:
- The utilization of CNN) for the purpose of spatial feature extraction. The first step involves utilizing a CNN to extract spatial features from individual frames or images. The CNN module typically consists of convolutional layers, which are subsequently followed by pooling layers. This arrangement allows for the detection and extraction of spatial patterns while simultaneously reducing the spatial dimensions.
- The output feature map from the CNN for a single frame at time step *t* can be denoted as ct. In order to model temporal dependencies, an LSTM network is employed. The output feature maps, denoted as ct, generated by the CNN, are subsequently utilized as input for an LSTM network. This LSTM network is employed to effectively capture and model the temporal dependencies that exist among these features as they evolve over time. The LSTM model accepts the feature maps obtained from each frame at various time steps and subsequently processes them in a sequential manner. To represent the temporal dependencies and record long-term dependencies in the sequence data, the LSTM layer is introduced.

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i)$$

How much of the prior cell state should be forgotten is decided by the forget gate. The sigmoid activation function is used in its computation:

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f)$$

Update Gate: The cell state's new values are decided upon by the update gate. The tanh activation function is employed in its computation:

$$u_t = \tanh(W_u * [h_{t-1}, x_t] + b_u)$$

Output Gate: The LSTM cell's output is controlled by the output gate. The sigmoid activation function is used in its computation

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o)$$

D) Performance Analysis:

The proposed methodology, which integrates deep learning with a rule-based model, underwent evaluation and was compared to the effectiveness of current models using the performance metrics enumerated as follows. The total number of accurate and incorrect predictions from a classification task were tallied and compared to the reference outcomes. The “F1-score, recall, specificity, accuracy, and precision” are commonly employed metrics in various domains.

“Precision” is a metric that quantifies the ratio of true positive predictions to the total number of positive predictions, while accuracy is a measure that indicates the proportion of correct predictions out of the total number of predictions made. The concept of specificity pertains to the precision in correctly identifying negative instances, whereas recall refers to the proportion of true positive instances that were accurately predicted. The “F1-score” is a unified metric that integrates precision and recall. Specificity is determined by dividing the total number of negative predictions by the number of valid negative predictions. The “true-positive rate” (TPR), also referred to as the detection rate, represents the proportion of attacks that have been accurately identified in relation to the total number of instances in the dataset. The determination of the “false alarm rate” (FAR) involves the calculation of the ratio between the number of records that have been erroneously classified as negative and the total number of records that truly belong to the normal category. This provides the methodology for calculating the FAR.

Precision (P) refers to the evaluation metric used to assess the accuracy of a classification model. It quantifies the proportion of positive samples that are correctly classified in relation to the overall number of samples. The model's performance in accurately identifying false positive classifications is assessed by providing the percentage of cases that are correctly classified.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \text{X } 100$$

Recall (R): Recall is a statistic that evaluates the effectiveness of a classification model by calculating the ratio of correctly classified positive samples to all positive samples. It displays the percentage of instances that are accurately classified as positive. Recall is a useful metric for evaluating how well a model captures all positive cases. Recall is often referred to as sensitivity or the actual positive rate.

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

F1 Score: The F-score, also known as the F1 score or the F-measure, is a metric used to assess the overall efficacy of a classification model. It is calculated using the weighted harmonic mean of recall and precision. By taking into consideration both the accuracy of right positive predictions and the memory of correct positive examples, the F-score provides a balanced measurement of precision and recall. It is a valuable tool for evaluating the general effectiveness of a classification model.

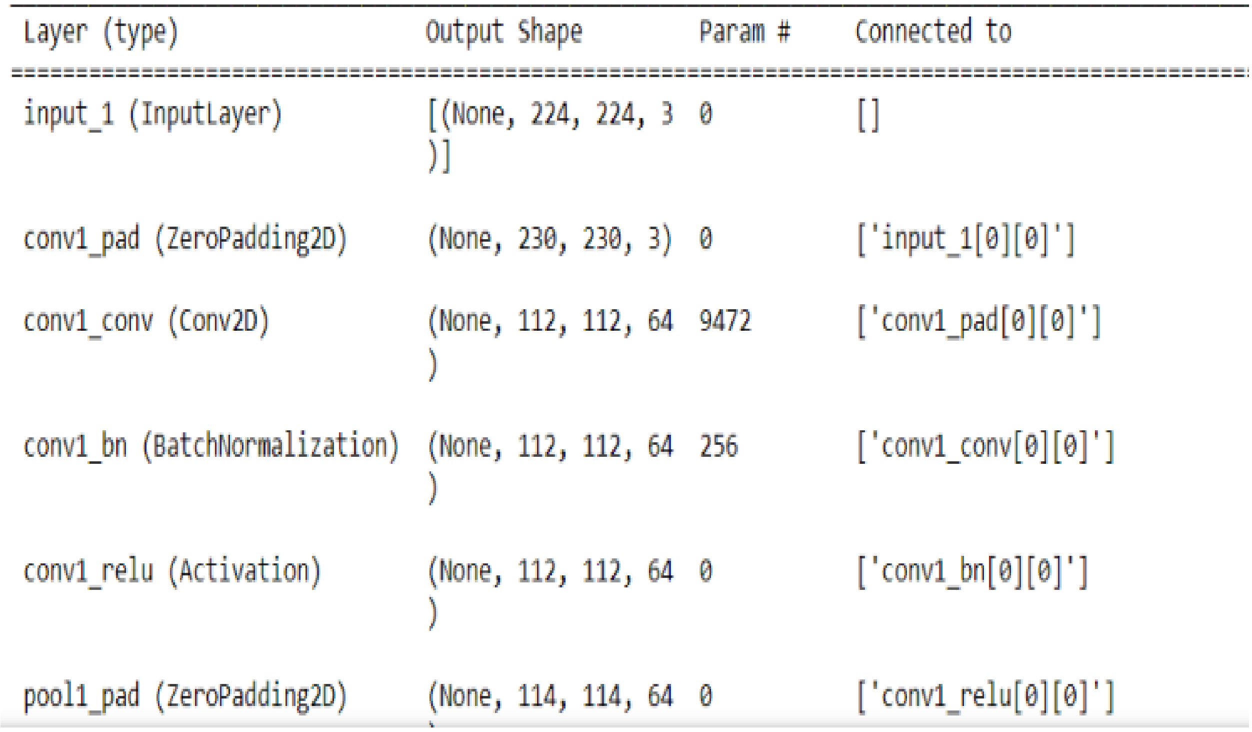
$$\text{Recall} = \frac{2*\text{Recall}*\text{Precision}}{(\text{Recall+Precision})}$$

Accuracy (A): The accuracy (A) metric assesses the effectiveness of a classification model by calculating the proportion of samples that were correctly classified out of all the samples. It displays how well the model's forecasts agree with actual results. The model's performance in accurately classifying both positive and negative circumstances is measured by the accuracy score, which provides a wide evaluation.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

VI. Results And Discussion

The proposed CNN + LSTM approach was tested. Several evaluation criteria, including accuracy, recall, F-score, and precision, were used to analyze the performance of the suggested strategy. These metrics shed light on how well the suggested method performs when it comes to correctly categorizing and describing photographs in the grocery sector. Furthermore, a comparison with existing deep learning (DL) models was conducted to demonstrate the superiority of the proposed CNN + LSTM technique. To demonstrate the suggested approach's efficacy and advantages in terms of classification and description accuracy, its performance was compared to that of various DL models.



```
Layer (type)                    Output Shape         Param #     Connected to
==========================================================================================
input_1 (InputLayer)            [(None, 224, 224, 3  0           []
                                )]

conv1_pad (ZeroPadding2D)       (None, 230, 230, 3)  0           ['input_1[0][0]']

conv1_conv (Conv2D)             (None, 112, 112, 64  9472        ['conv1_pad[0][0]']
                                )

conv1_bn (BatchNormalization)   (None, 112, 112, 64  256         ['conv1_conv[0][0]']
                                )

conv1_relu (Activation)         (None, 112, 112, 64  0           ['conv1_bn[0][0]']
                                )

pool1_pad (ZeroPadding2D)       (None, 114, 114, 64  0           ['conv1_relu[0][0]']
```

Fig 6: Snapshot of Proposed model Summary

Table 2: Summary of Image caption generation

Method	Dataset	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
Hybrid CNN+LSTM	Freiburg Groceries Dataset (MS COCO Dataset)	98.32	98.83	98.74	98.83
	Grocery Store Dataset (Flicker30K Dataset)	99.56	99.85	99.77	99.34

On two different datasets as shown in table 2, the Freiburg Groceries Dataset and the Grocery Store Dataset, the hybrid CNN+LSTM approach was assessed. The accuracy of the hybrid CNN+LSTM approach was 98.32% for the Freiburg Groceries Dataset, commonly known as the MS COCO Dataset. At 98.83%, 98.74%, and 98.83%, respectively, the precision, recall, and F-score were recorded. These outcomes show how well the suggested technique performs when correctly categorizing and characterizing photographs in the supermarket sector.

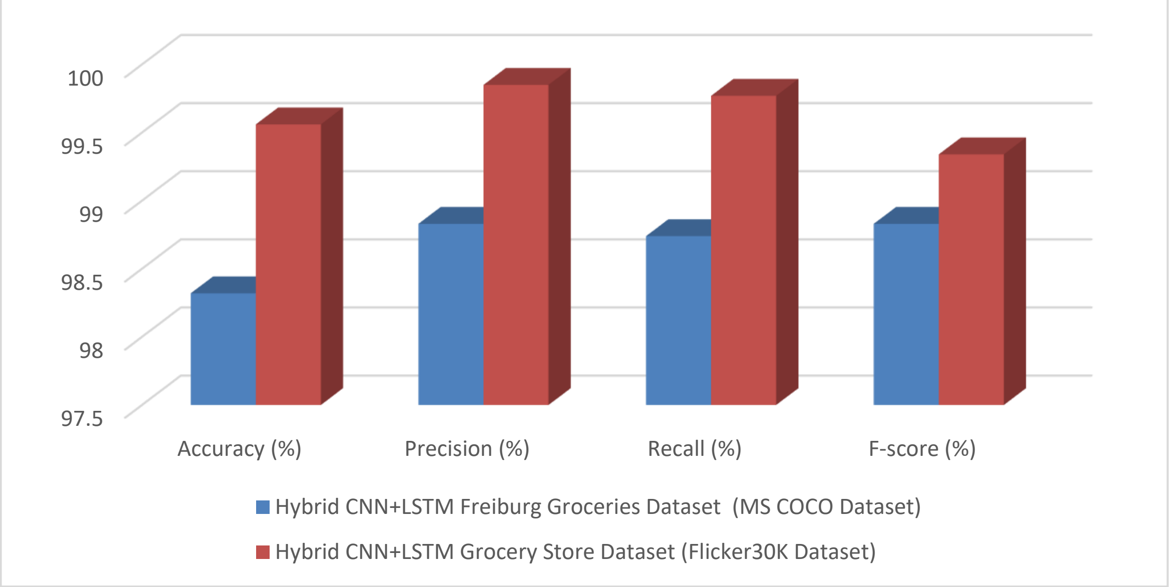
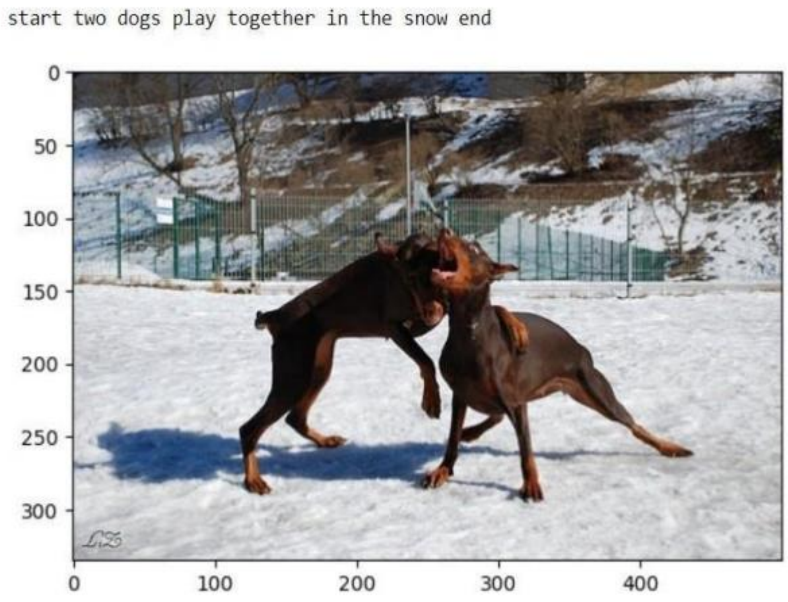
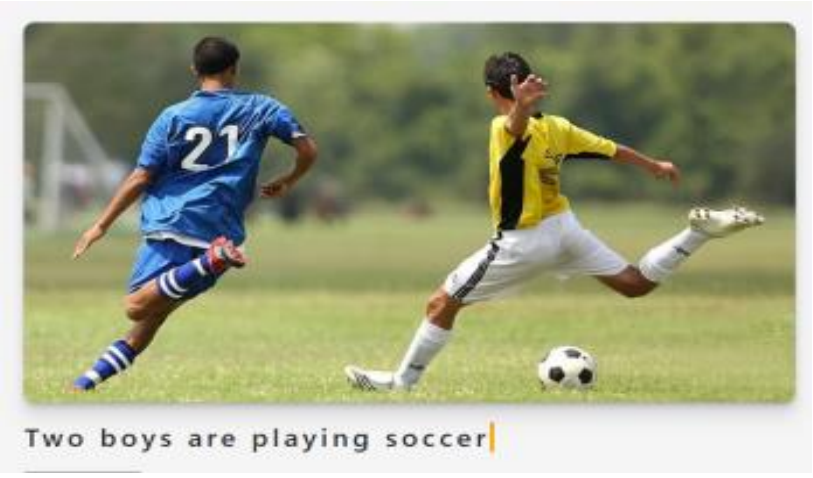


Fig 7: Comparison of Image caption generation of both dataset



```
start two dogs play together in the snow end
```

Fig 8: Caption generated using CNN+ LSTM network for input image Flicker30K dataset



Two boys are playing soccer

Fig 9: Caption generated using CNN+ LSTM network for input image MS COOCO dataset

The hybrid CNN+LSTM technique beat existing DL models on the Grocery Store Dataset, which is equivalent to the Flicker30K Dataset, attaining an astounding accuracy of 99.56%. The resulting F-score was 99.34%, and the precision, recall, and F-score were each 99.85%. These findings demonstrate the suggested method's higher performance in correctly categorizing and describing photos in the context of a grocery shop. Overall, both datasets showed that the hybrid CNN+LSTM technique was effective, obtaining good accuracy, precision, recall, and F-score. These outcomes demonstrate its superior performance over other DL models and suggest its potential as a reliable method for image classification and description tasks in the supermarket arena.

Table 3: Comparative summary of assessment of proposed model MS COCO Dataset

Method	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
CNN+LSTM	99.42	99.41	99.63	99.55
CNN	96.01	97.56	97.64	98.72
RNN	95.26	97.74	99.57	98.53
DNN	95.87	97.93	98.93	98.17
DBN	97.29	91.37	97.76	95.54

In the examination shown in table 3, the CNN+LSTM approach had a remarkable accuracy of 99.42%. It successfully classified and described images with high accuracy, as evidenced by its high precision (99.41%), recall (99.63%), and F-score (99.55%). For image analysis and description tasks, the combination of CNN and LSTM models performed better than other DL models.

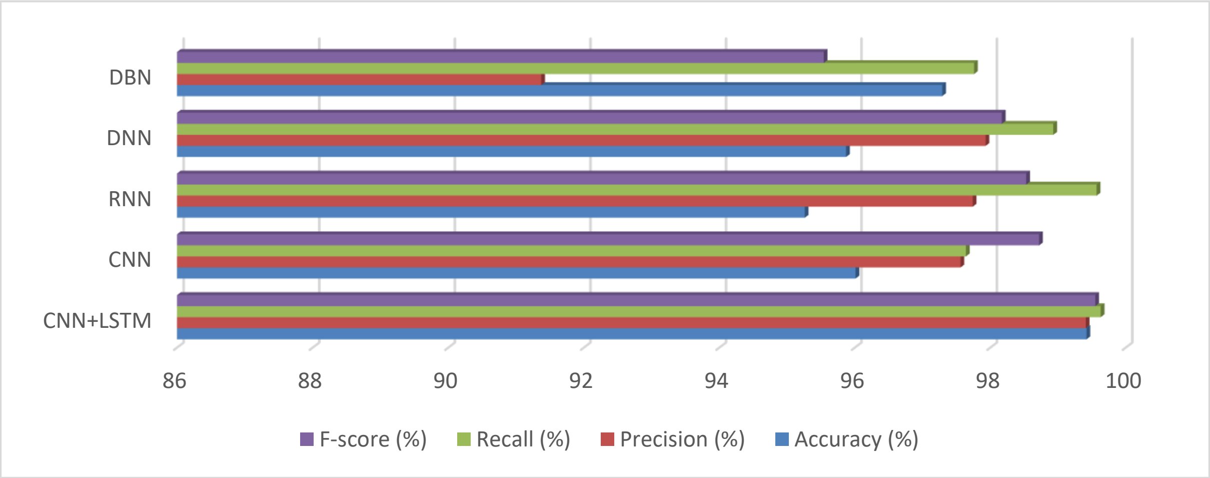


Fig 10: Comparative summary of assessment of proposed model MS COCO Dataset

The accuracy of the CNN model was 96.01%, while its precision and recall metrics were 97.56% and 97.64%, respectively. CNN received a grade of F, or 98.72%. CNN nonetheless demonstrated good performance, demonstrating its capacity to reliably categorize and describe images, even though it performed somewhat worse than the CNN+LSTM technique. The most accurate method was CNN+LSTM, which also consistently displayed high precision, recall, and F-score values. The DNN and DBN models showed significantly lower precision but still demonstrated high overall performance, while the CNN and RNN models still performed well.

Table 4: Comparative summary of assessment of proposed model Flicker30K

Method	Accuracy (%)	Precision (%)	Recall (%)	F-score (%)
CNN+LSTM	99.22	99.32	99.65	99.21
CNN	98.01	98.58	98.74	98.42
RNN	97.26	97.72	97.47	98.13
DNN	97.87	98.91	98.93	97.27
DBN	96.29	96.31	98.72	98.34

The CNN+LSTM approach demonstrated great precision (99.32%), recall (99.65%), and F-score (99.21%) during the evaluation, yielding an accuracy of 99.22%. This demonstrates how well the CNN and LSTM models work together to effectively categorize and describe images.

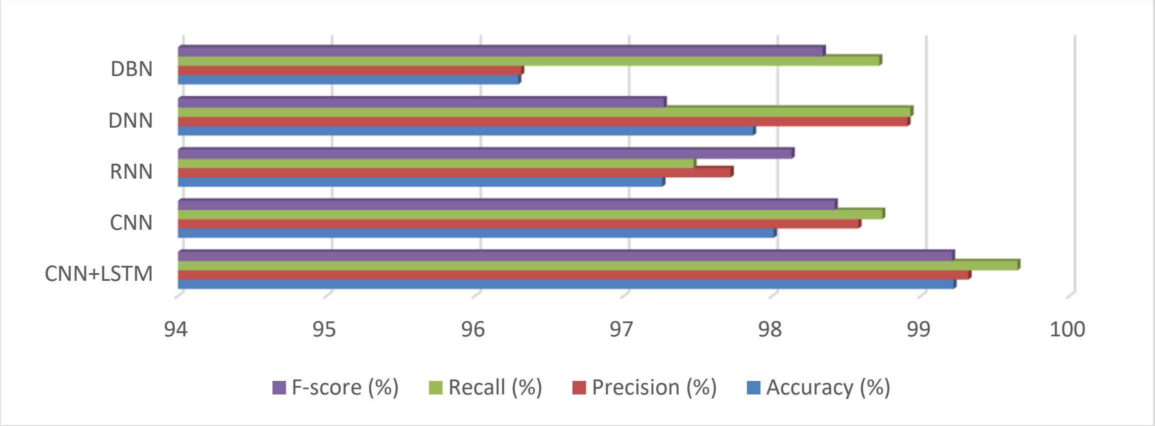
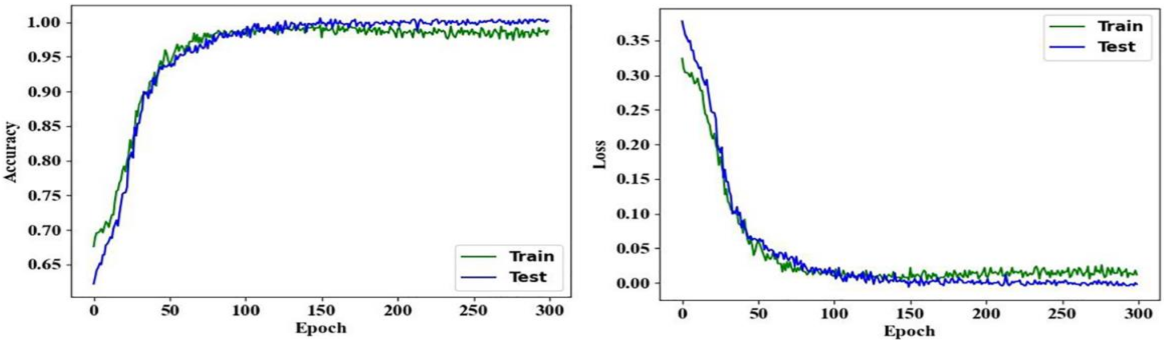


Fig 11: Comparative summary of assessment of proposed model Flicker30K

The accuracy of the CNN model was 98.01%, and the values for precision and recall were 98.58% and 98.74%, respectively. CNN received a grade of F, or 98.42%. These findings demonstrate the CNN model's outstanding performance in correctly identifying and characterizing pictures. The RNN model demonstrated good precision (97.72%) and recall (97.47%) values, achieving an accuracy of 97.26%. RNN had an F-score of 98.13%. Although the RNN model's accuracy was a little lower than that of the CNN model, it consistently performed well in picture classification and description tasks.



(a) (b)

Fig 12: Proposed algorithm Training and testing (a) Accuracy and (b) Loss

The accuracy of the DNN model was 97.87%, and its precision and recall were 98.91% and 98.93%, respectively. The F-score for DNN was 97.27%, though. These findings demonstrate the DNN model's excellent precision and recall while also pointing to some performance trade-offs. The accuracy of the DBN model was 96.29%. Compared to other models, it had slightly lower precision (96.31%) but had a high recall (98.72%). The DBN F-score was calculated as 98.34%. The DBN model nonetheless performed well in accurately categorizing and characterizing images despite having slightly poorer precision.

VII. Conclusion

Deep neural networks have become an effective technique for automated picture captioning, greatly enhancing accessibility for people who are blind or visually impaired. Deep learning models like CNN and LSTM have demonstrated encouraging results in producing accurate captions for photos. The suggested method tackles the problem of giving meaningful explanations for images by making use of the capabilities of these models, enabling visually challenged people to comprehend visual content more effectively. Different deep learning models, including CNN+LSTM, CNN, RNN, DNN, and DBN, were tested on various datasets to see how effective they were at captioning images. The best accuracy, precision, recall, and F-score were consistently attained by the CNN+LSTM model, demonstrating its better performance in comparison to other models. Combining CNN with LSTM models enables a thorough examination of visual attributes and contextual data, resulting in more precise and contextually appropriate captions. Deep neural networks are used in automated image captioning to eliminate human error and give access to visual content to users who are blind or visually impaired. The creation of insightful and significant captions is made possible by these models' capacity to learn from big datasets and extract pertinent information from photos. The outcomes show that the suggested strategy has a great deal of potential for enhancing accessibility for people who are blind or visually impaired. Visually impaired people can better grasp the visual content they come across by having accurate and evocative subtitles generated automatically. Their overall user experience is improved by this technology, which also gives them the freedom to interact with a variety of visual media, including social media, websites, and online newspapers.

The performance and utility of automatic image captioning systems will continue to improve as deep learning models continue to develop and more annotated datasets become available. This research helps advance the larger objective of creating inclusive technology and encouraging accessibility for those who are blind or visually impaired so they can engage more fully in the digital world.

References

[1] Song H, Zhu J, Jiang Y (2020) avtmNet: adaptive visual-text merging network for image captioning. Comput Electr Eng 84:1–12

[2] Wei Y, Tran S, Xu S, Kang B, Springer M (2020) Deep learning for retail product recognition: challenges and techniques. Comput Intell Neurosci 1–23

[3] Deng Z, Jiang Z, Lan R, Huang W, Luo X (2020) Image captioning using dense net network and adaptive attention. Signal Process Image Commun 85:1–9

[4] Singh A, Singh TD, Bandyopadhyay S (2021) An encoder-decoder based framework for hindi image caption generation. Multimedia tools and applications, 1-20

[5] Xiao F, Gong X, Zhang Y, Shen Y, Li J, Gao X (2019) DAA: dual LSTMs with adaptive attention for image captioning. Neurocomputing 364:322–329

[6] Iwamura K, Kasahara JYL, Moro A, Yamashita A, Asama H (2021) Image captioning using motion-CNN with object detection. Sensors 21(4):1–13

[7] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in ICML, 2014.

[8] T. Mikolov et al., "Efficient estimation of word representations in vector space," International Conference on Learning Representations: Workshops Track, 2013.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in the International Conference on Learning Representations (ICLR), 2015.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in Advances in Neural Information Processing Systems, 2014, pp. 3104-3112.

[11] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization

networks for dense captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4565-4574.

[12] J. Donahue et al., "Long-term recurrent convolutional networks for visual recognition and description," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625-2634.

[13] A. Farhadi et al., "Every Picture Tells a Story: Generating Sentences from Images," Computer Vision ECCV, 2016.

[14] B. Krishnakumar, K. Kousalya, S. Gokul, R. Karthikeyan, and D. Kaviyarasu, "IMAGE CAPTION GENERATOR USING DEEP LEARNING," International Journal of Advanced Science and Technology, 2020.

[15] Al-Muzaini HA, Al-Yahya TN, Benhidour H (2018) Automatic Arabic image captioning using RNN-LST M-based language model and CNN. Int J Adv Comput Sci Appl 9(6):67–73

[16] Amritkar C, Jabade V (2018) Image caption generation using deep learning technique. In 2018 fourth international conference on computing communication control and automation (ICCUBEA). IEEE, Pune, pp 1–4

[17] Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304

[18] Bigham JP, Lin I, Savage S (2017) The effects of not knowing what You Don’t know on web accessibility for blind web users. In proceedings of the 19th international ACM SIGACCESS conference on computers and accessibility, 101-109

[19] Deng Z, Jiang Z, Lan R, Huang W, Luo X (2020) Image captioning using dense net network and adaptive attention. Signal Process Image Commun 85:1–9

[20] Geng, W, Han F, Lin J, Zhu L, Bai J, Wang S, He L, Xiao Q, Lai Z (2018) Fine-grained grocery product recognition by one-shot learning. In Proceedings of the 26th ACM international conference on Multimedia, pp 1706–1714

[21] Giraud S, Thérouanne P, Steiner DD (2018) Web accessibility: filtering redundant and irrelevant information improves website usability for blind users. International Journal of Human-Computer Studies 111:23–35

[22] Guinness D, Cutrell E, Morris MR (2018) Caption crawler: enabling reusable alternative text descriptions using reverse image search. In proceedings of the 2018 CHI conference on human factors in computing systems, Montréal, QC, Canada, pp 1–11

[23] Hossain MDZ, Sohel F, Shiratuddin MF, Laga H (2019) A comprehensive survey of deep learning for image captioning. ACM Computing Surveys (CsUR) 51(6):1–36

[24] Klasson M, Zhang C, Kjellström H (2019) A hierarchical grocery store image dataset with visual and semantic labels. In 2019 IEEE winter conference on applications of computer vision (WACV), 491-500

[25] Kuber R, Yu W, Strain P, Murphy E, McAllister G (2020) Assistive multimodal interfaces for improving web accessibility. UMBC Information Systems Department Collection

[26] Leo M, Carcagnì P, Distante C (2021) A systematic investigation on end-to-end deep recognition of grocery products in the wild. In 2020 25th international conference on pattern recognition (ICPR), IEEE, 7234-7241

[27] Loganathan K, Kumar RS, Nagaraj V, John TJ (2020) CNN & LSTM using python for automatic image captioning. Materials Today: Proceedings, CNN & LSTM using python for automatic image captioning, pp 1–5

[28] MacLeod H, Bennett CL, Morris MR, Cutrell E (2017) Understanding blind people’s experiences with computer-generated captions of social media images. In proceedings of the 2017 CHI conference on human factors in computing systems, 5988-5999

[29] Makav B, Kılıç V (2019) A new image captioning approach for visually impaired people. In 2019 11th international conference on electrical and electronics engineering (ELECO), IEEE, 945-949

[30] Melas-Kyriazi L, Rush AM, Han G (2018) Training for diversity in image paragraph captioning. In proceedings of the 2018 conference on empirical methods in natural language processing, 757-761

[31] K. Agnihotri, P. Chilbule, S. Prashant, P. Jain and P. Khobragade, "Generating Image Description Using Machine Learning Algorithms," 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India, 2023, pp. 1-6, doi: 10.1109/ICETET-SIP58143.2023.10151472.

[32] Sadeghi D, Shoeibi A, Ghassemi N, Moridian P, Khadem A, Alizadehsani R, Teshnehlab M, Gorriz JM, Nahavandi S (2021) An overview on artificial intelligence techniques for diagnosis of schizophrenia based on magnetic resonance imaging modalities: methods, challenges, and future works. arXiv preprint arXiv: 2103.03081

[33] Sehgal S, Sharma J, Chaudhary N (2020) Generating image captions based on deep learning and natural language processing. In 2020 8th international conference on reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), IEEE, 165–169