Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

Conversational AI : LSTM-Based Image Recognition and Assistance for Blind Users

Ms. K. Sushmeena [1],
[1] *Assistant Professor, Department of CSE, Sri Bharathi Engineering College For Women, Pudukkottai*
sushmeena.nk@gmail.com

ABSTRACT: Image captioning is a challenging computer vision task that involves creating captivating text summaries for photos. This technology combines ideas from computer vision and natural language processing to comprehend an image's content and communicate it in a way that humans can understand. The fundamental necessity for inclusion and equitable access to knowledge is the reason why blind people require image captioning. This technology is crucial for providing descriptive textual information about the contents of images so that those with visual impairments can comprehend visual content that would otherwise be inaccessible to them.By allowing blind persons to independently investigate and understand the visual components of their environment, image captioning fosters a sense of autonomy and reduces need on sighted assistance. Many existing initiatives and systems that use image captioning technology are meeting the needs of the visually impaired. This study suggests a novel approach that builds an image captioning system using Convolutional Neural Network (CNN) techniques to increase accessibility for those with visual impairments.In order to help the blind understand visual content they frequently come across in their daily lives, the system attempts to provide meaningful and thorough descriptions of visuals. By utilizing CNNs, the model is able to extract and interpret pertinent characteristics from images, producing meaningful captions that are subsequently conveyed to users via assistive technologies like speech synthesis. By providing a possible solution to close the visual information gap and enable people with visual impairments to interact and navigate the visual world more effectively, the project tackles the critical need for inclusive technology.
effectively, the project tackles the critical need for inclusive technology.

INTRODUCTION

The goal of the computer science discipline of machine learning is to create algorithms that let computers learn from data without explicit programming. It is a branch of artificial intelligence that uses statistical methods to help machines learn and get better with time. Fundamentally, machine learning entails teaching computers on huge datasets so they can recognize trends and generate predictions. Because these predictions may be used to automate operations, optimize processes, or make choices, machine learning is a very valuable tool in many different industries.

The capacity of machine learning to manage sizable and intricate datasets is one of its main benefits. Machine learning algorithms can swiftly spot patterns and insights that human analysts might overlook due to the growing amount of data being generated in almost every business. Because of this, it's a useful tool for companies trying to enhance operations, spot growth prospects, and streamline decision-making. Machine learning's capacity for adaptation and long-term learning is another important benefit. Machine learning algorithms can increase the accuracy and refinement of their predictions by continuously analyzing new data. This enables companies to adjust to new trends, stay ahead of shifting market conditions, and make more educated decisions. Machine learning algorithms come in a variety of forms, such as reinforcement learning, supervised learning, and unsupervised learning. By using labeled datasets to train algorithms, supervised learning enables them to generate predictions based on the input data.In contrast, unsupervised learning uses unlabeled data to train algorithms that can recognize patterns and cluster data points according to commonalities. By teaching computers to base their judgments on input from

their surroundings, reinforcement learning enables them to learn by making mistakes. Even with all of its benefits, machine learning has drawbacks. The requirement for high-quality data is one of the field's main obstacles. Because machine learning algorithms are only as good as the data they are trained on, their predictions will also be erroneous if the data is inaccurate or lacking. The requirement for interpretability presents another difficulty for machine learning.

It might be challenging to comprehend how algorithms make their predictions as they get more complicated. In addition to making it harder to find and fix biases or mistakes in the algorithms, this can make it difficult for organizations to base their judgments on those forecasts. Notwithstanding these difficulties, machine learning has a lot of potential advantages. Machine learning algorithms help businesses remain ahead of the curve, make smarter decisions, and open up new avenues for growth and innovation by allowing machines to learn and get better over time.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

LITERATURE SURVEY

Dessi, Roberto, et al. [1] offered a straightforward fine-tuning technique to increase the discriminativeness of model-generated captions. The text generation component of a pre-trained captioner is optimized to assist a black-box textbased image retriever in selecting a target image from a variety of distracting images. We were able to get the system to function using the simple REINFORCE algorithm, and the task just required unannotated photos. We leave the investigation of increasingly complex reinforcement learning methods as a clear path forward. We present our findings with two captioners: BLIP, an encoderdecoder trained with multitask learning on webscale data, and ClipCap, a decoder-only model. Qualitatively, we observe that our discriminative finetuning technique recovers a more exact and clearly descriptive language even when finetuning the Conceptual Captions-trained captioner (which has learned to recreate the more abstract style of alttext descriptions). It is also evident why these more detailed captions, which eliminate the quirks of alttext, will translate more readily to other datasets than the original captioner's. Since ClipCap and Conceptual Captions showed the biggest difference between human and have been fixed by using VLP methods. Nevertheless, the majority of VLP techniques focus on comprehension tasks, while creation tasks like image captioning require additional capabilities.
discriminatively-generated captions and a noticeable asymmetry in retrieval vs. generation performance, we concentrated our investigation on these setups. Solomon Rodas et al. [2] incorporate an attention mechanism that concentrates on both visual and linguistic cues in an effort to overcome the shortcomings of earlier models. The attention method emphasizes high-level semantic elements that better define the visual content while enabling the model to extract only pertinent information. Furthermore, a BiGRU architecture is utilized, which records input data in both forward and backward directions. This improves the ability to capture both textual and visual information, which can reduce the gap between them and result in image captions that are more semantically rich. In order to extract visual characteristics that are subsequently projected into an LSTM model, the DNN uses a CNN as an encoder. Additionally, the authors provide the structurecontent neural language model (SC-NLM), a revolutionary decoder neural network language model that creates words by combining vector content and structure. By utilizing the advantages of both the content and the structural information, the method improves the image captioning's accuracy. To create semantically correct and fluid captions, researchers have improved the picture caption model in a number of ways. A new ensemble model including a 2-layer LSTM model and an Inception model was utilized to solve the image caption generating challenge.

Ghandi Taraneh et al. [3] offered to address the issue of image captioning, however there are still certain obstacles and unresolved issues. The quality of the datasets has a major impact on how well the supervised algorithms work. Nevertheless, datasets are unable to capture the real world regardless of how large they are, and the set of objects the detector is trained to discriminate determines the applicability of supervised approaches. However, datasets that include imagecaption pairings generally include more instances of a particular circumstance (for instance, "man riding a skateboard"). Instead of using actual observed objects, the model is unjustly biased by these instances in the training data to provide more captions that resemble those examples. Over-reliance on language priors in the supervised paradigm can also result in object-hallucination. Some of the issues with supervised techniques and object detector-based designs This need has been attempted to be met by several recent efforts discussed in this paper.

But more research and analysis are needed in this area. Additionally, detector-free designs are becoming more and more common. In these configurations, the detector is taken out for the endto-end visionlanguage pre-training. Afzal, Muhammad Kashif, et al. [4] research on Urdu-language generative picture captioning. In order to make Visio linguistic deep learning models efficient and suitable for datasets of low size, we propose a new dataset for Urdu picture captioning, annotation treatment, and generalization guidelines. We point out the limitations of conventional evaluation metrics in Urdu and demonstrate how semantics-driven methods like Bert-F1 and LASER could be suitable for assessing this assignment in Urdu. To improve the language model capability in the captioning, which is left for future work at this movement, one can employ transformer for the decoder component. In order to bridge the contextual gap between visual and linguistic components, the encoder is collaboratively tuned on top of the trained decoder by the picture to natural language connection. This enhances the visual component's compatibility with the language model by enabling the loss feedback to reach the image encoder.

Yehao Li, Jianjie Luo, et al. [5] explore the notion of enhancing linguistic coherence and visuallanguage alignment in the diffusion model for image captioning. We develop a novel semanticconditional diffusion process that enhances the diffusion model with extra semantic prior in order to validate our assertion. To further stabilize and enhance the diffusion process, a guided self-critical sequence training approach is developed. We provide empirical evidence that our solution outperforms the most advanced non-autoregressive methods. We are pleased to observe that, despite using the same Transformer encoder-decoder construction, our new diffusion model-based paradigm is able to outperform the competitive autoregressive method. The findings essentially show how promising the diffusion model is for image captioning.

PROPOSED SYSTEM

By creating an image captioning system especially for people with visual impairments, the suggested system seeks to satisfy the basic demand for inclusion and equitable access to information. This system will create descriptive written

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

summaries for images based on advances in computer vision and natural language processing, enabling blind people to understand visual content that would otherwise be incomprehensible to them. By offering insightful and thorough descriptions of images, the suggested system seeks to help blind users understand the visual content they come across on a regular basis. The system builds a chatbot that can react to picture inputs by utilizing deep learning, namely Long Short-Term Memory (LSTM) networks. A pre-trained Convolutional Neural Network (CNN) extracts features from the supplied image to start the image identification process. The LSTM network then processes these features and interprets them to produce a description of the scene or objects in the picture in natural language. This enables the system to provide descriptions such as "A street with cars and pedestrians" or "A person sitting at a table with a cup of coffee." The system also allows for vocal interaction; users may ask questions or seek additional information on the image, and the chatbot will reply in real time using text-to-speech (TTS). The system can keep context and produce logical responses thanks to the incorporation of LSTM, which makes the conversation seem more engaging and genuine. This device converts visual information into easily understood spoken information, giving blind persons a tool to navigate their environment more freely.

1.NATURAL LANGUAGE PROCESSING:

Natural language processing, or NLP, is the processing and analysis of human language using a range of methods. The following are several fundamental NLP algorithms, each having a distinct method for handling tasks such as sentiment analysis, machine translation, language generation, text classification, and more:

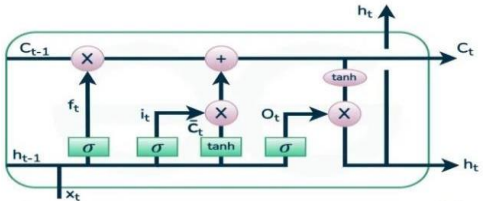
- **Tokenization :**is the process of breaking up text into discrete words, sentences, or phrases .
- **Stop Word Removal**: Eliminates common words that don't offer much sense, such as "the" and "is". Words are reduced to their base forms through stemming and lemmatization (e.g., "running" to "run").

2.LSTM ALGORITHM

- The input gate, forget gate, and output gate are the three gates that regulate the memory cell in LSTM architectures. The information that is added to, removed from, and output from the memory cell is determined by these gates.
- Information added to the memory cell is controlled by the input gate, information deleted from the memory cell is controlled by the forget gate, and information output from the memory cell is controlled by the output gate.
- The chain structure of the LSTM architecture comprises four neural networks and several memory units known as cells.
- The LSTM cell additionally features a memory cell that influences the cell's output at the current time step by storing data from earlier time steps.
- In order to handle and interpret sequential data over numerous time steps, the output of each LSTM cell is sent to the subsequent cell in the network.

3. DATASETS COLLECTION

High-quality image captioning datasets are crucial for training deep learning models, especially CNN and LSTM networks, to produce evocative and cohesive image captions. The development of the suggested system mostly depends on these datasets. These datasets usually include pictures together with captions that are authored by humans and use natural language to explain the images' contents. The model can produce insightful captions for previously viewed photos thanks to these datasets, which teach it to link visual cues with descriptions. Kaggle datasets are perfect for training image captioning models because they offer a wealth of varied images that are frequently accompanied by thorough captions. These datasets are easily accessible through downloaded links on Kaggle, and they are usually offered in CSV or JSON forms with linked or embedded image files.



4.TRAIN THE MODEL

We start by preprocessing the gathered datasets in order to train the image captioning model for the suggested system. To guarantee uniformity throughout the collection, images are downsized to a common size, such as 224x224 pixels, and their pixel values are normalized. To improve the dataset's diversity and lessen overfitting, data augmentation methods like flipping and random cropping can also be used. We tokenize the text for captions, turning each word into a distinct numeric ID, and use padding or truncation to guarantee that each caption has a set length. In order to map words to their integer representations, we also develop a glossary of terms that are used in the captions. An LSTM network and a convolutional neural network (CNN) make up the two primary parts of the model design. The CNN is used to extract feature vectors from images; it is frequently a pre-trained model, such as a sequential framework. These CNN models can capture crucial elements like edges, textures, and object shapes—all of which are essential for deciphering the photos' content—because they have already been trained on big datasets like ImageNet. The output from the penultimate layer is utilized as a feature vector that depicts the image after the CNN's final classification layer is eliminated. The LSTM network then receives this feature vector and uses it to create the caption. The input feature vector and previously predicted

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

words are used to train the LSTM to predict the next word in the caption, one word at a time. By understanding the connections between the extracted visual elements and the relevant words in the captions, the model is trained to produce descriptions of images that are both logical and contextually accurate.

5. IMAGE UPLOAD

Through a straightforward and intuitive interface, the suggested system's Image post feature enables users to post photographs with ease. Users can choose a picture from their device by clicking on a "Upload Image" button in a web or mobile application. The backend server receives the submitted image and processes it. To guarantee consistency and peak performance, the system then preprocesses the image by shrinking it to a standard size, like 224x224 pixels, and normalizing the pixel values. Following feature extraction from the CNN, the preprocessed image is sent to the LSTM network, which creates a descriptive caption based on the features. Through this process, the system is able to offer the user with meaningful descriptions of the image that can be read aloud to aid in their understanding of the visual information.

6. FEATURES EXTRACTION

A key phase in the image captioning process is feature extraction, which entails converting unprocessed picture data into a format that the model can comprehend and use to produce captions. Within the framework of the suggested system, a Convolutional Neural Network (CNN) is used to extract features. The CNN uses a number of layers to process the image, each of which is intended to identify various feature levels. While deeper layers identify more intricate patterns like forms, objects, or even people and backgrounds, early layers capture simpler elements like edges and textures. A feature vector, a condensed representation of the image's content, is produced by the model once the image has gone through the CNN layers. The Long Short-Term Memory (LSTM) network, which comes next, uses this feature vector—a numerical representation of the image's salient features—as input. The feature extraction stage is crucial for converting visual input into comprehensible English since the LSTM network uses this vector to produce a textual description of the image. Before producing a caption or response, feature extraction basically allows the algorithm to "understand" the image in terms of its essential visual elements.

7. VOICE DESCRIPTION

Voice Description is the feature that translates the generated textual descriptions of images into spoken words so that users who are blind or visually impaired can comprehend the image's content. This is a crucial part of the The caption is sent to a Text-to-Speech (TTS) engine after the system uses the CNN and LSTM networks to create one for the submitted image. The user is subsequently presented with natural-sounding speech that has been transformed from the text by the TTS engine.

SOFTWARE SPECIFICATION TENSORFLOW LIBARIES IN PYTHON

The Google Brain Team created the opensource machine learning framework TensorFlow. It is among the most widely used libraries for creating and refining deep neural networks and other machine learning models. TensorFlow makes it simple for developers to create sophisticated models for natural language processing, picture and audio recognition, and other applications. TensorFlow's capacity to manage intricate calculations and big datasets is one of its primary characteristics, which makes it appropriate for deep neural network training. Faster training times are made possible by the parallelization of computations across several CPUs or GPUs. Additionally, TensorFlow offers Keras, a high-level API that streamlines the model-building and training process.

Integration with other Python libraries and frameworks is made simple by TensorFlow's extensive collection of tools and libraries. Preparing data for training and analyzing model performance is made simple by its integrated support for data preprocessing and visualization. TensorFlow's ability to distribute models across multiple platforms, including as mobile devices and the web, is one of its main features. For deploying models on Android, iOS, and other mobile platforms, TensorFlow Lite is a mobile-optimized version of TensorFlow. TensorFlow.js is a JavaScript package that enables for training and deployment of models directly in the browser. To construct and train machine learning models, TensorFlow offers a variety of features and tools. TensorFlow's salient characteristics include:

COMPUTING BASED ON GRAPHS:

TensorFlow employs a computing model based on graphs, which enables effective computation
across a number of devices and CPUs/GPUs suggested
system since it enables audio feedback for user interaction.

Automatic Differentiation: TensorFlow's automatic differentiation feature makes it possible to compute gradients for backpropagation methods in an effective manner.
HIGH-LEVEL APIS: TensorFlow offers high-level APIs, like Keras, that let programmers create and train intricate models rapidly and with little code.

Preprocessing and Data Augmentation:
TensorFlow offers a variety of preprocessing and data augmentation techniques, such as data normalization and picture and text preprocessing
.
TensorFlow facilitates distributed training across a number of devices, CPUs, and GPUs, which enables quicker training periods and more economical resource usage.

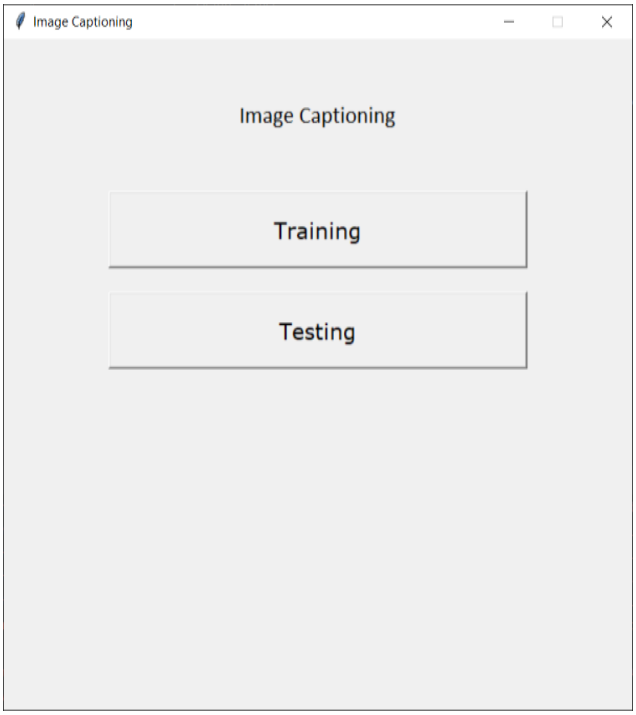
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

SYSTEM TESTING

Software testing is an inquiry carried out to notify stakeholders about the caliber of the service or product being tested. Software testing can also provide an objective, unbiased view of the software to assist the business to comprehend and understand the risks of software implementation. Test methods include, but are not restricted to, running an application or program in order to identify software faults (errors or other problems). One definition of software testing is the practice of confirming and confirming that a computer program, application, or product:

1. fulfills the specifications that influenced its creation and design;
2. functions as planned;
3. can be applied with identical features; and
4. meets stakeholder needs.

Depending on the testing methodology used, software testing can be applied at any stage of the development process. According to agile techniques, the majority of the test effort is ongoing, whereas traditionally it happens after the requirements have been established and the coding process is finished. Therefore, the chosen software development approach governs the test's technique. At different stages of the development process, the test effort will be concentrated by various software development models. Agile and other more recent development methods frequently use test-driven development and give developers more control over testing before it is sent to a formal team of testers. The majority of the test execution under a more conventional approach takes place following the completion of the coding phase and the definition of the requirements.



CONCLUSION

In summary, the seeing beyond vision system's use of Convolutional Neural Network (CNN)powered picture description is a major step forward in improving accessibility for the blind community. In addition to addressing the basic desire for inclusivity, this creative solution represents a positive step in using technology to empower and improve the lives of those with visual impairments. A more accessible and inclusive society for everyone can be our goal with the ongoing development and application of such inclusive technology.

Using deep learning methods such as Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for caption generation, the system can recognize objects and situations in photos with high accuracy. The generated descriptions can be pronounced aloud thanks to the inclusion of a Text-to-Speech (TTS) engine, which helps users comprehend and engage with the visual content around them. Through a conversational AI platform, this technology not only gives blind users more independence but also offers an accessible interface for interacting with the outside world. Future developments in the model, such better captioning accuracy and real-time interaction features, may increase its accuracy and usefulness even more, opening up new avenues for blind users to receive aid in their daily lives.

FUTURE ENHANCEMENTS

Human-written captions for a wider range of photographs might be collected through crowdsourced captioning, which would increase the system's accuracy and provide deeper descriptions. These captions could be utilized to improve the system's training and give the descriptions a more complex, human-like feel. In the future, it might be possible to identify sentimental or emotional content in photos. For instance, the system may determine whether the scene portrays a certain mood (such as serene, joyous, or melancholy) or whether a person in the picture is grinning, and then incorporate this information into the speech description.

REFERENCES

[1] Dessì, Roberto, et al. "Cross-domain image captioning with discriminative finetuning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[2] Solomon, Rodas, and Mesfin Abebe. "Amharic Language Image Captions Generation Using Hybridized Attention-Based Deep Neural Networks." Applied Computational Intelligence and Soft Computing 2023 (2023).
[3] Ghandi, Taraneh, Hamidreza Pourreza, and Hamidreza Mahyar. "Deep learning approaches on image captioning: A review." ACM Computing Surveys 56.3 (2023): 1-39.
[4] Afzal, Muhammad Kashif, et al. "Generative image captioning in Urdu using deep learning." Journal of Ambient Intelligence and Humanized Computing 14.6 (2023): 77197731.
[5] Luo, Jianjie, et al. "Semantic-conditional diffusion networks for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 07, NCIRTAI - 2026

[6] Mahalakshmi, P., and N. Sabiyath Fatima. "Summarization of text and image captioning in information retrieval using deep learning techniques." IEEE Access 10 (2022): 18289-18297.

[7] Sharma, Himanshu, et al. "Image captioning: a comprehensive survey." 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC). IEEE, 2020

[8] Muhammad Shah, Faisal, et al. "Bornon: Bengali image captioning with transformer-based deep learning approach." SN Computer Science 3 (2022): 1-16.

[9] Bhalekar, Madhuri, and Mangesh Bedekar. "D-CNN: a new model for generating image captions with text extraction using deep learning for visually challenged individuals." Engineering, Technology & Applied Science Research 12.2 (2022): 8366-8373.

[10] Sharma, Himanshu, and Anand Singh Jalal. "Incorporating external knowledge for image captioning using CNN and LSTM." Modern Physics Letters B 34.28 (2020): 2050315.