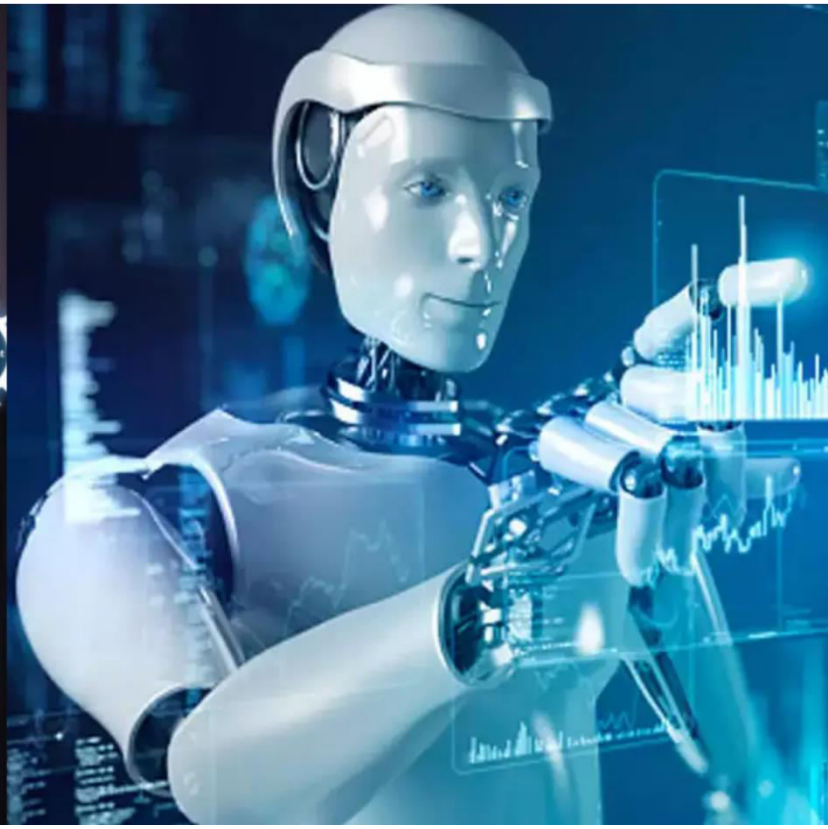
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 6, June 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406060|

Image Caption Generator with Multilingual Translation and Audio Narration

Koushik Roy[1], Vinay Kumar Aarya[2], Ravi Kumar[3]

MCA Student, Department of Computer Applications, Dr. B.C. Roy Engineering College, Durgapur, West Bengal, India[1,2,3]

ABSTRACT: This research introduces a comprehensive, deep learning-based architecture that performs automatic image caption generation and enriches the user experience with multilingual translation and text-to-speech (TTS) functionality. The system employs a pre-trained VGG16 convolutional neural network for extracting high-dimensional visual features from input images, which are then passed into a Long Short-Term Memory (LSTM) network for generating natural language captions that are both syntactically coherent and contextually relevant. To broaden accessibility and global usability, the generated captions are translated into multiple languages using the Google Translate API, allowing users to interact with the system in their native language. Additionally, the integration of the gTTS (Google Text-to-Speech) engine converts translated captions into audible speech, making the platform particularly useful for visually impaired individuals, language learners, and multilingual communication platforms.

The entire system is encapsulated within an intuitive web-based interface developed using Streamlit, which enables users to seamlessly upload images, select translation languages, view generated captions, and listen to speech outputs. The system is evaluated qualitatively on the Flickr8k dataset, demonstrating robust performance in terms of caption accuracy and semantic alignment with visual content. By unifying computer vision, natural language processing, and speech synthesis into a single framework, this project offers a practical, inclusive, and scalable solution for assistive technology, educational tools, and interactive AI applications.

KEYWORDS: Image Captioning, Deep Learning, Visual Feature Extraction, LSTM, VGG16, Google Translate API, gTTS, Multilingual NLP, Streamlit, Assistive Technology, Speech Synthesis, Flickr8k, Human-Computer Interaction, Neural Networks

I. INTRODUCTION

The ability to describe the contents of an image using natural language has long been a fundamental challenge in the field of artificial intelligence (AI). This task, known as image captioning, requires the integration of two core AI domains: computer vision, which interprets and analyzes visual content, and natural language processing (NLP), which generates coherent and grammatically accurate textual descriptions. The synergy between these fields enables machines to mimic human-like perception and communication.

In recent years, the advent of deep learning has significantly advanced image captioning capabilities. State-of-the-art models now leverage convolutional neural networks (CNNs) to extract high-level semantic features from images and recurrent neural networks (RNNs) or Long Short-Term Memory (LSTM) networks to generate fluent

and contextually relevant captions. Despite these advancements, most existing systems are limited in scope—they focus exclusively on generating captions in English and lack mechanisms for real-world deployment, accessibility, and multilingual usability.

This research addresses these limitations by developing a comprehensive, end-to-end image captioning system that not only generates high-quality English captions but also enhances accessibility and user engagement through multilingual translation and voice output capabilities. The proposed architecture integrates a pre-trained VGG16 model for visual feature extraction and a trained LSTM-based decoder for sentence generation. Additionally, the system employs Google Translate API for language translation and Google Text-to-Speech (gTTS) to convert the translated captions into natural-sounding speech.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406060|

To ensure ease of use and real-time interaction, a Streamlit-based user interface has been developed, enabling users to upload images, view automatically generated captions, select target languages, and listen to the spoken output. This inclusive design is particularly beneficial for visually impaired users, language learners, and educational platforms, thereby expanding the practical applicability of image captioning systems in diverse, real-world contexts.

II. LITERATURE SURVEY

Image captioning has gained significant attention in recent years due to advancements in deep learning, particularly the integration of vision and language models. Most existing architectures follow an encoder-decoder paradigm, where Convolutional Neural Networks (CNNs) are used for image feature extraction, and Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks are utilized to generate descriptive textual outputs.

Dr. S. Pasupathy proposed a CNN-LSTM-based caption generation system evaluated using the BLEU score metric, demonstrating effective word-to-image alignment [1]. Prof. S. Sankareswari et al. enhanced this model by integrating attention mechanisms, allowing the decoder to focus on different image regions during caption generation. The system also incorporated advanced evaluation metrics such as METEOR and CIDEr for improved semantic accuracy [2]. Aryan Chauhan et al. conducted a comparative study of architectures like Long-term Recurrent Convolutional Networks (LRCN) and Neural Image Caption (NIC) models, emphasizing the role of datasets such as Flickr8k, Flickr30k, and MS COCO in training captioning systems [3].

While these methods have shown substantial progress in generating accurate captions, they primarily focus on monolingual, text-based outputs. None of the reviewed works integrated audio narration or multilingual support, which are crucial for accessibility, especially for visually impaired or non-English-speaking users. This paper builds upon the existing literature by proposing a novel system that not only generates image captions but also translates them into multiple languages and converts them to speech using gTTS and Google Translate, all within a real-time web-based application.

III. METHODOLOGY

A. System Architecture

The proposed system consists of Some major components: The proposed system is a modular, deep learning-based pipeline that processes an image input and delivers a multilingual, speech-enabled caption through the following sequential steps:

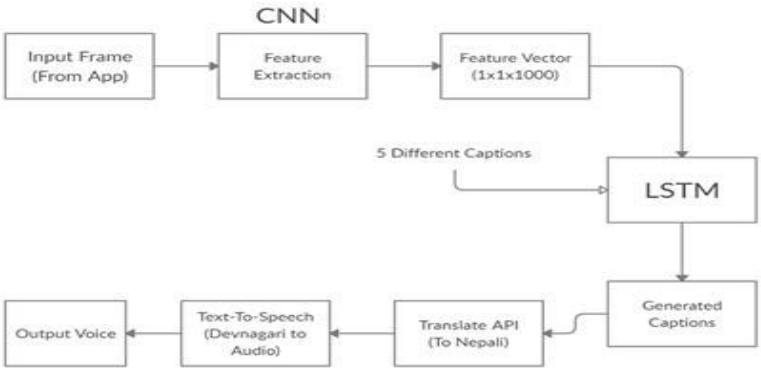


Figure 1: Architecture of the Image Caption Generator System

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406060|

Feature Extraction (VGG16)

- Input images are resized and preprocessed to fit the input requirements of the VGG16 model.
- The VGG16 model, pre-trained on the ImageNet dataset, is used to extract a 4096-dimensional feature vector from the image.
- This vector represents the semantic contents and structure of the visual data.

Caption Generation (LSTM + Tokenizer)

- The image feature vector is passed into a trained LSTM-based captioning model.
- Captions are pre-tokenized using Keras Tokenizer and padded to ensure uniform sequence lengths.
- The model sequentially predicts the next word in the sentence until an <end> token is reached, producing a meaningful English caption.

Translation Module (Google Translate API)

- The English caption is sent to the Google Translate API for multilingual translation.
- Users can choose from supported languages such as Bengali, Hindi, French, etc.
- The translated caption is stored for both display and speech synthesis.

Text-to-Speech (gTTS)

- The translated caption is passed to the Google Text-to-Speech (gTTS) engine.
- gTTS converts the text into an MP3 audio file, enabling voice playback.
- This makes the system highly beneficial for visually impaired users and non-readers.

User Interface (Streamlit)

- The entire workflow is wrapped inside an interactive Streamlit web application.
- Users can:
  - Upload an image
  - View the generated English caption
  - Select a language for translation
  - Listen to the caption in audio form

B. Dataset

- Flickr8k dataset is used.
- 8000 images with 5 captions each.
- Captions are preprocessed (lowercase, punctuation removed, start/end tokens added).
- Vocabulary tokenized using Keras Tokenizer.

C. Training Strategy

- Feature vectors stored using pickle.
- Captions encoded to sequences.
- Model trained using cross-entropy loss with teacher forcing.
- Checkpoints used to save best-performing weights.

IV. EXPERIMENTAL RESULTS

Sample Output

The figure 2 below demonstrates a complete run of the image caption generation system. A user uploads an image through the Streamlit interface, and the system generates a caption using the trained LSTM model. The caption is then translated into a user-selected language (in this case, Hindi), and finally converted to speech using the gTTS library. The interface also allows users to play.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406060|

In the given example, the uploaded image shows a dog chasing a soccer ball. The system successfully generates the caption: **"a dog chasing a soccer ball"**

The translated Hindi caption is:
"एक कुत्ता एक फुटबॉल गेंद का पीछा करते हुए"
The audio version of the caption can be played directly from the interface using the embedded audio player.

Figure 2: Streamlit interface displaying generated caption, translated output, and audio player.

Image-to-Caption with Voice Narration: Step-by-Step Process
When a user uploads an image—for example, one depicting "a dog playing with a soccer ball"—the system follows a structured pipeline to generate a **textual caption** and its **audible voice output**. The process can be broken down into the following modules:

Step 1: Image Upload and Preprocessing

- **Component Used**: Streamlit
- **Functionality**: The user uploads an image via a web interface.

- **Process**:
  - Image is resized to the expected input shape (e.g., 224x224x3).
  - Image pixels are normalized and reshaped as required by the CNN model.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406060|

- **Output**: Preprocessed image array (NumPy format).

Step 2: Feature Extraction

- **Component Used**: VGG16 (Pre-trained CNN model from Keras)
- **Functionality**: Extract high-level visual features from the image.
- **Process**:
  - Image is passed through the VGG16 model (excluding the top classification layer).
  - The output is a 4096-dimensional feature vector representing the semantic content of the image.

  features = model predict(preprocessed image)
- **Output**: Feature vector (used as input to the caption generator).

Step 3: Caption Generation

- **Components Used**: LSTM Decoder, Keras Tokenizer
- **Functionality**: Generate a natural language caption based on the image feature.
- **Process**:
  1. The model begins captioning with a special token: <start>.
  2. The LSTM decoder predicts the next word in sequence using:
     - The image feature vector.
     - Previously predicted words.
  3. This loop continues until an <end> token is predicted.

  Caption = "a dog playing with a soccer ball"
- **Output**: English text caption describing the image.

Step 4: Language Translation (Optional)

- **Component Used**: googletrans (Google Translate API)
- **Functionality**: Translate the caption into a user-selected language.
- **Process**:
  from googletrans import Translator
  translator = Translator ()
  translated = translator.translate(caption, dest ='hi') # Example: Hindi
- **Output**: Translated text: "एक कुत्ता एक फुटबॉल गेंद का पीछा करते हुए" (Hindi)

Step 5: Text-to-Speech Conversion

- **Component Used**: gTTS (Google Text-to-Speech)

- **Functionality**: Convert the translated (or English) caption into speech.
- **Process**:
  from gtts import gTTS
  tts = gTTS(text=translated.text, lang='hi') # or lang='en' for English
  tts.save("caption_audio.mp3")
- **Output**: An MP3 audio file containing the spoken caption.

Step 6: Audio Playback via Web Interface

- **Component Used**: Streamlit Audio Player
- **Functionality**: Allow users to listen to the generated voice narration.
- **Process**:
  st.audio("caption_audio.mp3")
- **Output**: Embedded audio player with play controls.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 6, June 2025

|DOI: 10.15680/IJIRSET.2025.1406060|

Summary of Module Functions and Outputs:

Step	Component	Function	Output
1	Streamlit	Upload and preprocess image	Resized image array
2	VGG16	Extract image features	4096-d feature vector
3	LSTM + Tokenizer	Generate English caption	"a dog playing with a soccer ball"
4	Google Translate	Translate caption to selected language	e.g., Hindi: " एक कुत्ता एक फुटबॉल गेंद का पीछा करते हुए "
5	gTTS	Convert text to speech	MP3 audio file
6	Streamlit UI	Display caption, play audio	Caption text and audio output on interface

V. CONCLUSION

This paper presents an accessible, multilingual, and interactive image caption generator that effectively combines computer vision and natural language processing techniques. By integrating VGG16 for feature extraction, LSTM for sequence generation, gTTS for speech synthesis, and Google Translate for multilingual support, the system not only generates high-quality captions but also makes them available in audio and different languages. This solution has practical applications for the visually impaired, educational platforms, and multilingual content creation. Future work will involve integrating attention mechanisms, adopting transformer models, and deploying the application as a cloud service for wider reach.

REFERENCES

1. Dr. S. Pasupathy, “Image Caption Generator using CNN and LSTM,” International Journal for Multidisciplinary Research (IJFMR), vol. 5, no. 2, pp. 1–6, Mar-Apr 2023.
2. Prof. S. Sankareswari et al., “Image Caption Generator Using Deep Learning,” International Journal of Creative Research Thoughts (IJCRT), vol. 11, no. 10, Oct. 2023.
3. Aryan Chauhan et al., “Research Paper on Image Caption Generator Using Deep Learning,” Journal of Emerging Technologies and Innovative Research(JETIR), vol. 10, no. 4, pp. c802–c806, 2023.
4. Keras Applications Documentation - VGG16. [Online]. Available: https://keras.io/api/applications/vgg/
5. gTTS: Google Text-to-Speech Python Library. [Online]. Available: https://pypi.org/project/gTTS/
6. googletrans: Free Google Translate API for Python. [Online]. Available: https://pypi.org/project/googletrans/
7. Flickr8k Dataset. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details