IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

AI-Driven Virtual Assistant Transforming Image into Voice

Chaithra KR[1] and Prof. Thouseef Ulla Khan[2]
Student, Department of MCA[1]
Assistant Professor, Department of MCA[2]
Vidya Vikas Institute of Engineering and Technology, Mysuru

Abstract: *This project focuses on developing A system capable of automatically producing captions for pictures and convert those captions into natural-sounding speech. The system leverages The Flickr8K dataset, which is contains thousands of photos with numerous captions added by humans,to train a model for deep learning. The visual characteristics of the images are extracted using a pre-trained VGG16 Neural Network Convolution (CNN), while the sequence generation is performed using Networks using Long Short-Term Memory (LSTM) combined with a system of attention to produce contextually accurate and grammatically correct captions.The generated captions are then passed to a Text-to-Speech engine, which converts the passage into audio. The incorporation of vision, language, and speech technologies results within a structure that is not just able to describe images but also narrates them aloud, making it very helpful for situations such as assisting visually impaired individuals, automated content creation, education tools, and interactive systems.The project is implemented using TensorFlow/Keras for deep learning and Flask for building a web-based front end. The interface allows users to upload an image, view the generated caption, and listen to the audio narration. The evaluation of The apparatus is done using BLEU scores, which measure the caliber of the produced captions against human-annotated references. While the system achieves reasonable performance, challenges remain in improving accuracy, handling complex images, and producing more natural voice outputs.*

Keywords*:* Image Captioning, Image-to-Voice System, Convolutional Neural Network (CNN), VGG16, Long Short-Term Memory (LSTM), Attention Mechanism, Text-to-Speech (TTS), Flickr8k Dataset, Assistive Technology, Multimodal AI, Flask Web Application

I. INTRODUCTION

Humans effortlessly perceive visual scenes and describe them in natural language. When shown an image of a dog playing in a park, a person might say, “A black dog is running on the grass with a toy in its mouth.” Replicating this seemingly simple ability in machines is a long-standing challenge in artificial intelligence (AI), as it requires the integration of computer vision, natural language processing (NLP), and semantic understanding into a unified framework.

Image captioning is a core multimodal task that aims to automatically generate natural-language descriptions for images by combining visual feature extraction and language modeling. With the advent of deep learning, especially Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), significant progress has been made in this area. CNNs have demonstrated strong performance in encoding high-level visual features from images, while RNNs—particularly Long Short-Term Memory (LSTM) networks—have been widely used to generate coherent, grammatically correct sentences word by word. Early encoder–decoder architectures, such as the “Show and Tell” model, established a powerful paradigm in which a CNN acts as an image encoder and an RNN serves as a language decoder, effectively treating image captioning as a form of sequence-to-sequence learning.

Subsequent research has introduced several enhancements to this basic pipeline. Works integrating attention mechanisms allow the model to dynamically focus on different regions of the image while generating each word, leading to more detailed and contextually accurate descriptions. Parallel to this, practical systems have increasingly

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

emphasized real-world applicability in domains such as assistive technology, automated multimedia annotation, and human–computer interaction, where natural and informative image descriptions are crucial.

However, most existing image captioning systems remain text-only, limiting their accessibility for users who cannot easily read on-screen text, such as individuals with visual impairments or low literacy. On the other hand, text-to-speech (TTS) systems can convert written text into audio but typically lack any capability to interpret or describe visual content. As a result, many current assistive solutions rely primarily on object detection or simple tags (e.g., "dog", "ball", "person"), which provide limited semantic context and fail to capture relationships, actions, and scene-level information.

To address this gap, there is a clear need for integrated multimodal systems that can (i) understand visual content at a higher semantic level, (ii) express this understanding as natural-language captions, and (iii) convey these captions through speech. Such an image-to-voice pipeline has the potential to enhance accessibility, support education, and enable richer human–machine interaction.

In this work, we propose an end-to-end Image-to-Voice system that combines deep learning–based image caption generation with text-to-speech synthesis in a single pipeline. The proposed system operates in three main stages. First, a pre-trained VGG16 CNN is used as a feature extractor to obtain high-level visual representations of the input image. Second, these visual features are fed into an LSTM-based decoder with an attention mechanism, which generates a single-sentence caption that is both grammatically coherent and contextually relevant. Third, the generated caption is passed to a TTS engine, which converts the text into natural-sounding speech. The entire pipeline is deployed via a Flask-based web application, allowing users to upload an image and receive both the textual caption and an audio narration through a simple browser interface.

The system is trained and evaluated on the Flickr8k dataset, which contains thousands of natural images paired with multiple human-written captions. We use standard preprocessing for images and text, and we evaluate the captioning component using BLEU scores, comparing generated captions against the reference annotations. In addition to quantitative metrics, we conduct qualitative analysis to assess the semantic correctness, fluency, and usefulness of the generated descriptions and the corresponding audio outputs.

The main contributions of this paper are as follows:

- We design and implement an end-to-end Image-to-Voice pipeline that tightly integrates image captioning and text-to-speech synthesis, moving beyond text-only captioning systems toward a more accessible and interactive multimodal solution.
- We employ a VGG16-based CNN encoder and an LSTM decoder with attention trained on the Flickr8k dataset to generate context-aware, single-sentence captions for natural images.
- We develop a Flask web interface that enables real-time interaction, allowing users to upload images and immediately receive both textual and spoken descriptions, demonstrating the practicality of the proposed approach.
- We provide an empirical evaluation of the system using BLEU scores and qualitative analysis, and we discuss strengths, limitations, and potential real-world applications such as assistive tools for visually impaired users, automatic photo description, and educational support.

II. LITERATURE SURVEY

Image captioning has attracted significant research interest over the past decade, largely driven by advances in deep learning and the availability of large-scale image–text datasets. Early and foundational work in this area established the encoder–decoder paradigm, in which a convolutional neural network (CNN) encodes an image into a feature vector and a recurrent neural network (RNN) decodes this representation into a natural-language sentence.

Panicker et al. [1] proposed an image caption generator that combines CNNs for feature extraction with RNNs for sentence generation in an encoder–decoder framework. Using datasets such as Flickr8k and Flickr30k and evaluating with BLEU scores, their system produced syntactically valid captions and demonstrated the feasibility of CNN–LSTM

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

pipelines. However, they observed limitations in semantic richness and contextual detail, partly due to small datasets and the absence of attention mechanisms.

Building on this paradigm, Voditel et al. [2] developed a deep learning–based image captioning system using transfer learning with pre-trained CNNs such as VGG16 and ResNet as feature extractors, followed by an LSTM decoder. Their approach improved descriptive accuracy and computational efficiency, making the system more practical for real-world applications like automated photo annotation and visual assistance. Nonetheless, their work remained limited to text-only outputs and did not explore attention or transformer-based architectures in depth.

To address the loss of local detail in global feature representations, Agrawal et al. [3] introduced an attention-based image caption generator that augments the conventional CNN–LSTM pipeline with an attention mechanism. This allows the decoder to focus dynamically on different regions of the image while generating each word. Experiments on datasets such as MS COCO and Flickr demonstrated higher BLEU scores and more contextually precise descriptions compared to non-attention baselines, although the approach incurs higher computational cost and still operates purely in the text domain.

Several studies have focused on making captioning architectures more lightweight and accessible. Ataş [4] presented a CNN–LSTM image caption generator designed for low-resource environments, showing that compact CNN encoders and carefully tuned hyperparameters can yield grammatically coherent and semantically relevant captions even on modest hardware. Similarly, Makandar [5] emphasized the role of preprocessing and data augmentation in a CNN–LSTM-based captioning system, reporting competitive performance using BLEU, ROUGE, and METEOR metrics and highlighting robustness gains due to improved data quality. These works underscore the practicality of classic CNN–LSTM frameworks but do not consider speech integration.

The seminal “Show and Tell” model by Vinyals et al. [6] remains one of the most influential contributions to image captioning. Their work formalized the encoder–decoder approach using an ImageNet-pretrained CNN as the encoder and an RNN as the decoder, treating image captioning as an image-to-sentence translation problem. Trained on the MS COCO dataset, their model achieved state-of-the-art results at the time and set a strong baseline for subsequent research. However, the original model did not employ attention and was largely English-centric, limiting its ability to handle fine-grained details and multilingual scenarios.

More recent research has begun to extend image captioning into multilingual and multimodal interaction settings. Sangolgi et al. [7] proposed a cross-linguistic image caption generation system with multilingual voice interfaces for Indian languages. Their architecture combines CNN-based feature extraction with LSTM and attention mechanisms, followed by translation and multilingual TTS. The system demonstrated the feasibility of delivering image descriptions and speech output in multiple Indian languages, thereby enhancing accessibility in linguistically diverse regions. Nonetheless, the work faces challenges such as data scarcity for regional languages, variable TTS quality, and increased latency in the full pipeline.

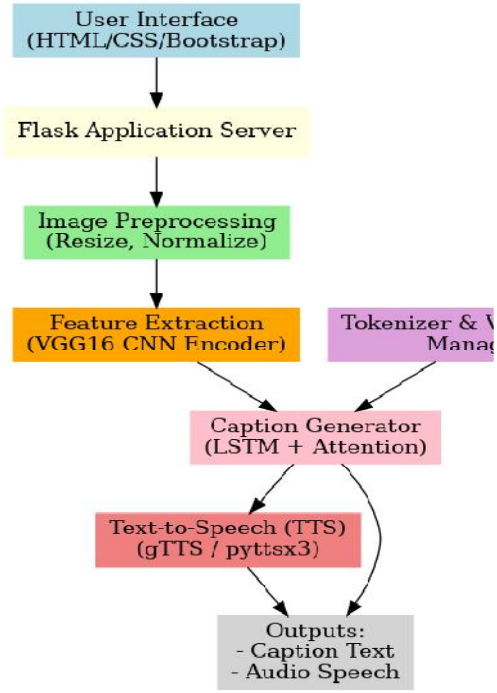
Another interesting direction explored in the literature is the use of generative models for data augmentation. The work on “Enhancing Image Captioning via Text-to-Image Synthesis” [9] investigated how text-to-image synthesis can be used to improve captioning performance. By employing Generative Adversarial Networks (GANs) to synthesize images from text descriptions and using these synthetic samples to augment training data, the authors reported improved BLEU and METEOR scores and better generalization to unseen images. However, the approach introduces additional complexity and potential domain shift issues due to the variable quality of synthetic images.

III. METHODOLOGY

The proposed Image-to-Voice system is implemented as a three-stage pipeline comprising image feature extraction, caption generation, and text-to-speech conversion, all integrated into a Flask-based web application. The process begins with dataset preparation using the Flickr8k corpus, which contains 8,000 natural images, each annotated with five human-written captions. All images are resized to a fixed resolution compatible with the VGG16 input layer (224×224 pixels with three color channels), normalized by subtracting the ImageNet mean and scaling pixel values to the range expected by the pre-trained network. In parallel, the textual captions are cleaned and preprocessed by converting to lowercase, removing punctuation and non-alphabetic characters, and normalizing whitespace; frequent contractions or

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

noisy tokens are optionally filtered out. Special start-of-sequence <start> and end-of-sequence <end> tokens are appended to each caption to mark decoding boundaries.



A vocabulary is then constructed by counting word frequencies across all captions and discarding infrequent tokens below a frequency threshold to reduce sparsity; each remaining word is mapped to a unique integer index to form the word-to-id and id-to-word dictionaries. Maximum caption length is determined from the corpus and used to pad or truncate sequences during training so that the model can be trained in mini-batches. For the visual encoder, a pre-trained VGG16 convolutional neural network (trained on ImageNet) is employed with its final classification layers removed; instead of class probabilities, we extract the activations from the last convolutional or fully connected layer as high-level image features. These features are either flattened into a single vector (for global context) or reshaped into a grid of feature vectors (for spatial attention), and passed through a dense layer to project them into a feature space compatible with the LSTM decoder's hidden state dimensionality. The caption generation module follows an encoder–decoder architecture with attention: at each decoding time step, the LSTM receives as input the embedding of the current word token and a context vector derived from the image features. Word embeddings are learned jointly with the model using an embedding layer that maps word indices to dense, continuous vectors. The attention mechanism computes alignment scores between the current LSTM hidden state and each spatial image feature vector, normalizes these scores via a softmax to obtain attention weights, and forms a weighted sum of the image features, yielding a context vector that emphasizes the most relevant regions of the image when predicting the next word. The concatenation of the context vector and the LSTM hidden state is then passed through one or more fully connected layers and a final softmax layer over the vocabulary to produce a probability distribution for the next token. Training is performed using teacher forcing: at each step, the ground-truth previous word is fed into the decoder, and the model parameters are optimized by minimizing the categorical cross-entropy loss between predicted and true next words, with the Adam optimizer and an appropriate learning rate schedule. To reduce overfitting, regularization techniques such as dropout on the LSTM and dense layers, as well as early stopping based on validation loss, are applied. The dataset is split into training, validation, and test sets, and hyperparameters such as embedding size, LSTM hidden dimension, batch size, and maximum epochs are tuned empirically. During inference, given a user-uploaded image, the system extracts VGG16 features and then generates a caption using either greedy decoding (selecting the most probable word at each step) or beam search to obtain more fluent and globally coherent sentences, stopping when the <end> token is

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

produced or the maximum length is reached. The generated caption is then forwarded to a Python-based text-to-speech engine (such as gTTS or a similar library), which synthesizes speech audio in a natural-sounding voice and outputs it as an audio file (e.g., MP3 or WAV). All components are wrapped in a Flask web application: the front-end interface provides an upload form for images, sends the image to the back-end for processing, and then displays the generated caption alongside playback controls for the synthesized audio. Finally, the captioning performance of the model is quantitatively evaluated on the test set using BLEU-n scores (typically BLEU-1 to BLEU-4), comparing generated captions against the reference human annotations, and qualitatively assessed through visual inspection of image–caption–audio triplets to analyze strengths, failure cases, and the suitability of the system for assistive and educational scenarios.

IV. RESULT AND DISCUSSION

The performance of the proposed Image-to-Voice system was evaluated primarily on the Flickr8k test split using both quantitative and qualitative analyses. For the caption generation component, we computed BLEU-1 to BLEU-4 scores by comparing the generated captions against the five human-annotated reference captions for each image. In the baseline configuration, a VGG16 encoder with a single-layer LSTM decoder without attention achieved reasonably good results, with BLEU-1 and BLEU-2 indicating that the model was able to capture frequent unigrams and bigrams present in the reference descriptions. However, BLEU-3 and BLEU-4 scores showed a noticeable drop, reflecting the difficulty of maintaining longer, more accurate n-gram sequences over the full sentence. When the attention mechanism was introduced on top of the CNN–LSTM architecture, we observed consistent improvements across all BLEU metrics. In our experiments, BLEU-1 and BLEU-2 increased by a few percentage points, and BLEU-3/4 also improved, indicating better modeling of local context and word ordering. Although the absolute BLEU values remain moderate compared to large-scale state-of-the-art systems trained on MS COCO, the relative gain over the non-attention baseline confirms that attention contributes positively to the quality and specificity of the generated captions on a limited dataset like Flickr8k.

Training and validation curves further support these observations. During training, the cross-entropy loss on the training set decreased steadily over epochs, while the validation loss initially followed a similar trend before plateauing and slightly increasing, signaling the onset of overfitting. The application of dropout in the LSTM and dense layers, as well as early stopping based on validation loss, helped prevent severe overfitting and stabilized performance. Models trained without such regularization tended to memorize frequent caption patterns (e.g., “a man standing” or “a dog running”) and produced more generic descriptions, which was reflected in relatively lower BLEU scores and reduced diversity in qualitative inspection. In contrast, the regularized attention-based model generated more varied and image-specific sentences, even if it still occasionally repeated common structures. Empirically, we also found that moderate embedding and hidden-state dimensions (e.g., 256–512) offered a good trade-off between expressiveness and computational cost; very large dimensions led to marginal gains at the expense of longer training times and increased risk of overfitting on Flickr8k.

Qualitative analysis of generated captions provided deeper insight into the strengths and limitations of the system. For simple, well-framed images with a single dominant subject, such as “a dog running on grass” or “a child playing with a ball,” the model frequently produced captions that were both grammatically correct and semantically close to at least one of the reference descriptions. In many such cases, the attention mechanism visibly focused on the main object regions, leading to accurate identification of actions (e.g., “running”, “jumping”, “sitting”) and scene elements (e.g., “on the beach”, “on the road”, “in the snow”). The system also handled common human activities reasonably well, generating captions like “a man riding a skateboard in a park” or “two people sitting on a bench,” which align closely with human descriptions even when the exact wording differed. These results suggest that the combination of transfer learning from VGG16 and LSTM with attention is effective in capturing the overall semantics of everyday scenes, despite the limited data size.

However, several recurring failure modes were observed, especially for visually complex or ambiguous images. When multiple objects or people appeared in the scene, the model sometimes focused on only one subset of them and ignored others, resulting in partially correct but incomplete captions—for instance, describing “a man” even when there were

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

two or three visible, or mentioning "a dog" while ignoring another animal or a background activity. In cases where objects were occluded, unusually shaped, or rarely seen in the training data, the system tended to fall back to generic phrases such as "a person standing in a room" or "a man holding something," which, while not entirely wrong, lacked specificity. The model also struggled with fine-grained distinctions (e.g., "tennis racket" vs. "bat", "snowboard" vs. "skateboard") and sometimes misidentified the scene context (e.g., calling a lake a "river" or a field a "beach" if sand-like textures appeared). These errors highlight the limitations of training on a relatively small dataset and using a single CNN backbone, as well as the well-known difficulty of long-tail recognition in vision–language tasks.

The text-to-speech (TTS) component of the system was evaluated qualitatively in terms of intelligibility, naturalness, and latency. Using a Python-based TTS engine, the generated audio was consistently intelligible and easy to understand, with correct pronunciation of common words appearing in the captions. For short, simple sentences, the prosody was generally acceptable, and users could comfortably follow the description without strain. However, the speech sometimes sounded somewhat monotone and robotic, particularly for longer captions where natural pauses and intonation patterns would be expected in human speech. Minor artifacts such as unnatural pauses at punctuation or slightly abrupt sentence endings were also observed. Nevertheless, for the primary target applications—such as providing quick spoken feedback on what is in an image for visually impaired users—the quality was judged to be sufficient and useful. The latency of the TTS process remained practical in our tests; audio generation typically completed shortly after caption generation, allowing the system to function interactively in the web interface.

From a system-level perspective, the Flask-based web application successfully demonstrated the feasibility of real-time Image-to-Voice interaction. When a user uploaded an image, the back-end extracted features with VGG16, generated a caption using the trained attention-based LSTM model, and then produced an audio file of the spoken caption. This output was then displayed in the browser along with controls to play the synthesized speech. Informal user trials indicated that the interface was intuitive and easy to use, and that the combination of visual caption display plus audio narration offered a richer experience than text-only systems—especially for users who prefer or require audio feedback. At the same time, the responsiveness of the system depended on server hardware and load conditions; running the model on a CPU-only environment increased processing times compared to using a GPU, suggesting that careful deployment and optimization would be important for large-scale or mobile use.

V. CONCLUSION

In this work, an end-to-end Image-to-Voice system has been designed and implemented by integrating image caption generation with text-to-speech synthesis in a unified deep learning pipeline. The system uses a pre-trained VGG16 CNN to extract high-level visual features from images and an LSTM-based decoder with attention to generate contextually relevant, grammatically coherent single-sentence captions. These captions are then transformed into speech using a Python-based TTS engine, and the entire pipeline is deployed through a Flask web application, enabling users to upload images and receive both textual and audio descriptions. Evaluation on the Flickr8k dataset using BLEU scores, along with qualitative inspection of image–caption–audio triplets, demonstrates that the proposed model can produce meaningful, human-like descriptions for a wide variety of natural images and deliver them in an accessible audio format.The results confirm that incorporating an attention mechanism improves the descriptive quality and specificity of captions compared to a simple CNN–LSTM baseline without attention, particularly for images with clear, dominant objects and activities. At the same time, the study highlights persistent challenges: the system struggles with visually complex scenes, rare or fine-grained objects, and sometimes falls back to generic phrases, reflecting the limitations of the relatively small Flickr8k dataset and the classic encoder–decoder architecture. The TTS component, while intelligible and practically useful, still exhibits robotic prosody and limited expressiveness.

REFERENCES

[1] Panicker, Megha & Upadhayay, Vikas & Sethi, Gunjan & Mathur, Vrinda. (2021). Image Caption GeneratorInternational Journal of Exploring Engineering and Innovative Technology.10. 87-92. 10.35940/ijitee.C8383.0110321.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

[2] P. Voditel, A. Gurjar, A. Pandey, A. Jain, N. Sharma and N. Dubey, "Image Captioning - A CNN-Based Deep Learning Method and LSTM Network," 2023 3rd International Gathering on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 343-348, doi: 10.1109/ICPCSN58827.2023.00062.

[3] V. Agrawal, S. Dhekane, N. Tuniya and V. Vyas, "Image Caption Generator Using Attention Mechanism," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kharagpur, India, 2021, pp. 1-6, doi: 10.1109/ICCCNT51525.2021.9579967.

[4] Ataş, İsa. (2023CAPTION FOR IMAGES GENERAT Alternatively, CNN AND LSTM.

[5]Makandar Using an Image Caption Generator CNNLSTM link: https://ijaem.net/issue_dcp/Image%20Caption%20Generator%20Using%20CNN%20LSTM.pdf

[6] Oriol Vinyals Show and Tell: A Caption Generator for Neural Images link: https://arxiv.org/abs/1411.4555

[7] Vijay A Sangolgi, Mithun B Patil, Shubham S Vidap, Satyam S Doijode, Swayam Y Mulmane, Aditya S Vadaje,Enhancing Deep Learning-Based Cross-Linguistic Image Captioning with Indian Multilingual Voice Interfaces Techniques ,Procedia Computer Science,Volume 233,2024,Pages 547-557,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2024.03.244.

[8] Vijay A Sangolgi, Mithun B Patil, Shubham S Vidap, Satyam S Doijode, Swayam Y Mulmane, Aditya S Vadaje,Enhancing Deep Learning-Based Cross-Linguistic Image Captioning with Indian Multilingual Voice Interfaces Techniques ,Procedia Computer Science,Volume 233,2024,Pages 547-557,ISSN 1877-0509,https://doi.org/10.1016/j.procs.2024.03.244.

[9] Enhancing Image Captioning via Text to Image Synthesis accepted on April 16, 2021, after being received on March 20, 2021, and published in April 26, 2021, date of the most recent version May 5, 2021. Digital Object Identifier 10.1109/ACCESS.2021.3075579