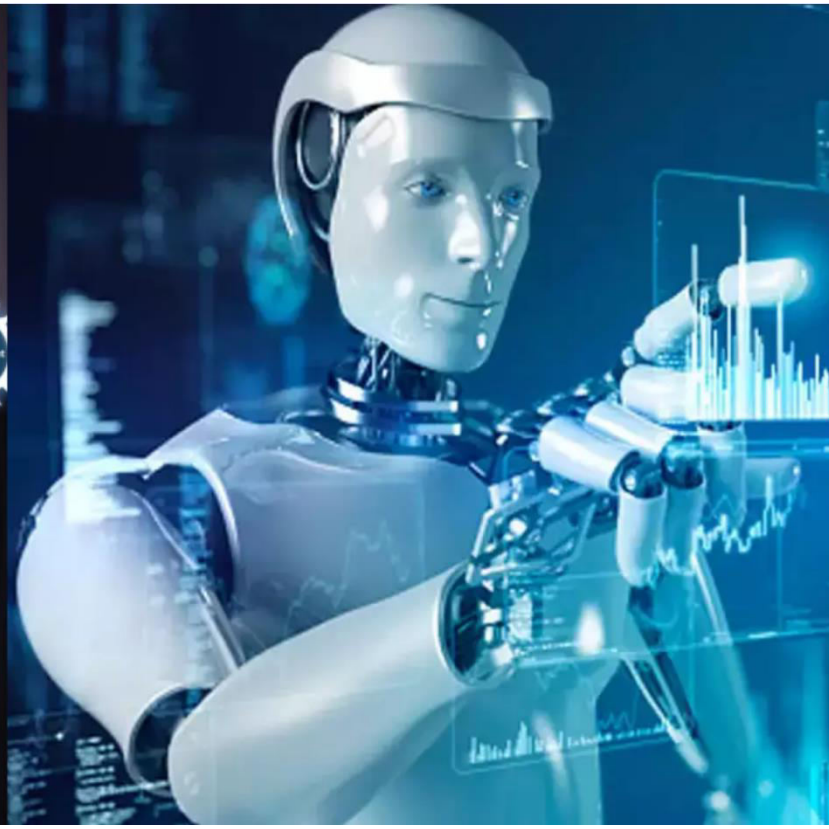
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

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 15, Issue 2, February 2026

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 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502087

From Pixels to Sentences: Automated Image Captioning with CNNs RNNs

Sumit Kumar, Srivastav, Himanshu Verma, Himanshu Verma, Himanshi Sharma, Ishwari Bhalerao, Kanani Hetvi

Ass. Professor, Department of Computer Science & Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, Gujarat, India

Department of Computer Science & Engineering, Parul University, Vadodara, Gujarat, India

Department of Computer Science & Engineering, Parul University, Vadodara, Gujarat, India

Department of Computer Science & Engineering, Parul University, Vadodara, Gujarat, India

Department of Computer Science & Engineering, Parul University, Vadodara, Gujarat, India

Department of Computer Science & Engineering, Parul University, Vadodara, Gujarat, India

ABSTRACT: The ability to automatically describe visual content through natural language is an exciting area in artificial intelligence research. Our work tackles this challenge by developing a complex neural architecture that converts visual information into clear textual descriptions. Our approach involves two stages: first, we use the strong feature extraction abilities of InceptionV3, a well-known convolutional neural network, to analyze the visual elements in uploaded images. The extracted visual features then feed into our custom language generation pipeline, which is based on a Gated Recurrent Unit (GRU) architecture. What sets our implementation apart is the addition of a spatial attention module that allows selective focus on different areas of the image while forming captions. This attention-driven method reflects how humans process visuals, emphasizing certain regions as we describe a scene. To demonstrate the practical use of our research, we created an easy-to-use web-based platform with the Streamlit framework. This interactive system lets users upload photos and get instant captions, along with audio narration using integrated speech synthesis technology.

KEYWORDS: Image Captioning, deep Learning, Attention Mechanism, Encoder-Decoder, Inceptionv3, GRU, Computer Vi- sion Natural Language Processing.

I. INTRODUCTION

Every day, countless photos are uploaded to the internet. However, computers interpret These images are just arrays of numbers. The goal of image captioning is to enable machines to interpret images in a manner akin to human perception and to articulate findings using natural language. This field is very valuable for many different applications. It supports accessibility by helping vision-impaired users com- prehend web-based imagery, facilitates content-based photo searches, and enables automated categorization of extensive image collections [11]. Our research addresses this challenge through a deep learning framework utilizing the established encoder-decoder architecture [1]. The encoding component leverages InceptionV3, a robust pre-trained convolutional neu- ral network, to extract and interpret visual information from images [6]. The decoding component employs a recurrent neu- ral network that converts this visual analysis into descriptive text. Our primary objective focused on creating a system that is capable of processing any input image and producing clear, precise captions.

Despite recent advancements, image captioning remains a challenging task due to the inherent complexity of visual scenes and the ambiguity in natural language descriptions. Variations in lighting, perspective, object occlusion, and scene composition can significantly affect feature extraction, while language generation must maintain grammatical coherence and contextual relevance. Addressing these challenges requires models that not only recognize objects accurately but also understand relationships, actions, and context within the scene. By combining CNN-based visual

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 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502087

encoding with GRU-based sequence generation and attention mechanisms, our approach aims to create captions that are both semantically meaning- ful and contextually precise, improving upon the limitations observed in previous works.

II. LITERATURE REVIEW

Today's image captioning systems mostly rely on the encoder-decoder architecture. Researchers adapted this approach from machine translation and implemented it in two key stages [1]. First, the encoder processes the input image. Studies have shown that using pre-trained networks like VGG [4] or InceptionV3 [6] significantly improves performance. These networks have already been trained on large datasets like ImageNet, which enables them to extract complex visual features and understand object-level information. The decoder then converts these extracted features into readable sentences. Recurrent Neural Networks (RNNs) are commonly used for this purpose because they naturally handle sequential data, capturing the flow of words in sentences [4].

Early RNN-based decoders faced challenges with long-term dependencies, often forgetting important information in longer sequences. Long Short-Term Memory (LSTM) networks [5] and Gated Recurrent Units (GRUs) were introduced to address this problem. They use gating mechanisms to decide which in- formation to retain or discard, significantly improving caption generation quality over long sequences.

A major advancement in image captioning has been the integration of attention mechanisms [3]. Instead of relying on a single compressed representation of the image, attention allows the model to focus on relevant regions dynamically while generating each word. For example, when describing "a dog catching a frisbee," the attention model can shift from focusing on the dog to the frisbee as the sentence progresses. This selective focus mimics human visual perception and improves both syntactic and semantic accuracy.

Further research has explored combining spatial and tem- poral attention [7], allowing models not only to focus on specific areas of an image but also to consider the sequence of generated words for better coherence. Transformer-based architectures have recently been introduced [2], which replace RNNs with self-attention layers, enabling parallel processing of sequences and capturing long-range dependencies more effectively. Such approaches have shown promise in produc- ing richer, more descriptive captions, especially for complex scenes with multiple objects.

Some works also highlight the importance of dataset di- versity. Datasets like MS COCO [9] and Visual Genome [10] provide dense annotations and a wide variety of scenes, which help models generalize better and reduce bias toward common captions. However, dataset imbalance still results in models producing generic descriptions in certain cases, emphasizing the need for more diverse and comprehensive datasets.

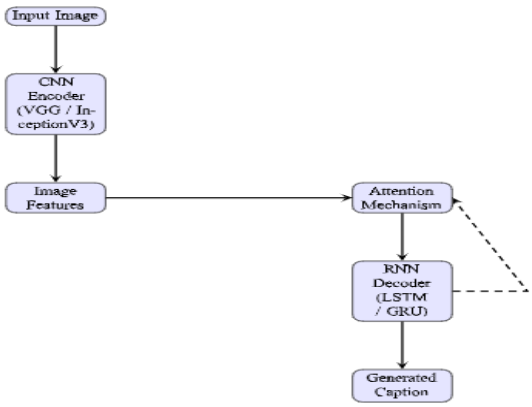


Fig. 1. Encoder-Decoder Architecture with Attention for Image Captioning. The CNN extracts image features which the RNN decoder uses with attention to generate textual captions.

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 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502087

III. PROPOSED SYSTEM ARCHITECTURE

Our system is built upon the encoder-decoder framework enhanced with an attention mechanism to generate high-quality captions. The workflow begins with InceptionV3, a pre-trained convolutional neural network, which acts as the feature extractor. Instead of using only the final classification layer, we extract feature maps from the last convolutional layer to retain spatial information about objects and regions in the image [6]. These rich feature representations are then passed into the encoder module, which projects them into a lower- dimensional embedding space suitable for the decoder.

The decoder is implemented using a Gated Recurrent Unit (GRU), which generates captions sequentially, one word at a time. The attention mechanism guides the GRU by focusing on specific regions of the image during each step of word generation, mimicking human visual attention [3]. This ap- proach ensures that the generated captions are contextually relevant and semantically accurate, rather than relying solely on a compressed global image representation.

To provide a clear overview of the system, Table I summa- rizes each module, its input, output, and function. This tabular representation allows readers to quickly grasp the workflow and understand the role of each component.

TABLE I
SUMMARY OF PROPOSED IMAGE CAPTIONING SYSTEM MODULES

Module	Input	Output
Image Preprocessing	Raw Image	Resized & Normalized Imag
Feature Extraction (CNN)	Preprocessed Image	Feature Map
Encoder	CNN Feature Map	Embeddings
Attention Mechanism	Embeddings + Hidden State	Context Vector
GRU Decoder	Context + Previous Words	Next Word
Audio Synthesis (Optional)	Generated Caption	Audio File

This structured architecture enables efficient end-to-end training while maintaining interpretability of the model's decision-making. By combining convolutional feature extrac- tion, attention-guided decoding, and sequential GRU predic- tion, the system achieves a balance between capturing visual information and generating natural language descriptions. Fu- ture extensions could include Transformer-based decoders or multimodal embeddings to further enhance caption richness [2].

IV. IMPLEMENTATION

The proposed image captioning system is implemented in Python using TensorFlow and Keras [4], [6]. The system consists of three main components: a CNN-based feature extractor, an attention-guided GRU decoder, and a web-based interface for real-time interaction [3], [5].

A. Data Preparation

All captions were tokenized into individual words to build a comprehensive vocabulary. Special start-of-sequence (<start>) and end-of-sequence (<end>) tokens were added to each caption so the model can identify the beginning and end of each sentence. Images were resized to a standard di- mension and normalized, ensuring consistency during feature extraction. Data augmentation techniques, such as random flipping and rotation, were also applied to enhance the model's robustness.

B. System Components

The InceptionV3 CNN is used to extract meaningful fea- tures from input images. These features are then processed by the GRU decoder with attention, which sequentially generates captions in natural language. The attention mechanism

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 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502087

allows the decoder to focus on the most relevant regions of the image while producing each word, improving both semantic accuracy and descriptive richness. During training, teacher forcing was employed to accelerate convergence and stabilize the learning process.

C. Deployment

The system is deployed as a web application using Streamlit [12]. Users can upload images, generate captions in real time, and listen to them through integrated Google Text-to-Speech, enhancing accessibility for visually impaired users. The inter- face is designed to be simple, intuitive, and responsive, making it practical for both demonstration and evaluation purposes. Logging and error handling modules ensure reliability during user interaction, while a caching mechanism reduces redundant computations, improving efficiency.

D. Hyperparameters

TABLE II
KEY HYPERPARAMETERS FOR TRAINING

Hyperparameter	Value
Embedding Dimension	256
GRU Units	512
Batch Size	64
LearninRate	0.001
Optimizer	Adam
Dropout	0.5
Max Caption Length	30

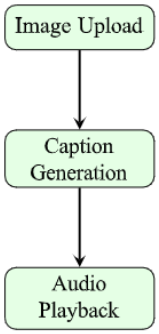


Fig. 2. Streamli Interface Workflow: Users upload Images, captions are Generated, using the GRU decoder with attention, and audio is played via TTS.

V. RESULTS AND DISCUSSION

We evaluated our image captioning model using the Stream- lit web interface. The captions generated were syntactically correct and semantically meaningful. For instance, for images like “a man throws a frisbee” or “a family enjoying a pic- nic,” the system produced accurate and contextually relevant descriptions. The integrated text-to-speech module performed reliably, converting captions into clear audio output, thereby enhancing accessibility for visually impaired users.

The model’s performance relies on the complementary roles of its core components. The CNN encoder extracts detailed image features [4], while the GRU decoder constructs coherent sentences. This combination ensures reliable caption genera- tion across diverse test images. Nevertheless, some limitations were observed. In certain instances, the model generated generic captions such as “a group of people are standing,” which, although correct, lacked descriptive richness. This phenomenon is primarily due to dataset bias: simpler captions dominate large datasets like MS COCO [9], leading the model to prefer common, safe phrases over detailed descriptions [8]. To quantitatively evaluate the model, we computed BLEU scores for a set of test images. BLEU (Bilingual Evaluation Understudy) measures the overlap between generated captions and reference human annotations, providing a standard metric for image captioning performance. Table III summarizes the

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 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502087

results:

TABLE III
EVALUATION METRICS (BLEU SCORES)

Metric	Image 1	Image 2	Image 3
BLEU-1	0.78	0.72	0.81
BLEU-2	0.64	0.59	0.69
BLEU-3	0.52	0.47	0.55
BLEU-4	0.39	0.35	0.41

E. User Interface

The BLEU scores indicate that the model captures relevant visual and linguistic information, particularly in short n-gram matches (BLEU-1 and BLEU-2). Lower scores in BLEU-3 and BLEU-4 suggest that longer sequences or highly specific phrases are more challenging, reflecting the model's tendency to favor general but syntactically correct expressions.

Overall, the experiments confirm that our encoder-decoder framework with attention and GRU decoding effectively gener- ates meaningful captions. It successfully balances grammatical correctness and semantic relevance, while BLEU evaluation provides measurable evidence of its performance. Future work Could include training on, larger and more diverse datasets to reduce generic captions, exploring Transformer-based decoders for even richer sequence generation.

V. CONCLUSION AND FUTURE WORK

A. Conclusion

We successfully created a comprehensive image captioning system for this project that can produce insightful descriptions for a variety of images. Our model can focus on significant areas of an image while generating grammatically and semantically accurate captions by fusing a CNN-based encoder with a GRU decoder that has been improved by attention mechanisms. In order to make the system accessible to visually impaired users and demonstrate its usefulness in the real world, it was implemented as a web application with an integrated text-to-speech feature. Scores from the BLEU evaluation verified that the captions produced by our model are accurate and consistent across a range of situations. Overall, this work shows that careful integration of deep learning architectures us Attention mechanisms Can create systems which bridge the gap between visual perception and natural language, offering significant potential for Applications in accessibility and multimedia understanding [1], [4], [6].

B. Future Work

There are a number of encouraging avenues to improve this system even more. Reducing generic captions and increasing descriptive accuracy would be possible by adding more complex and varied images to the training dataset [10]. A deeper comprehension of long-range dependencies in text and the creation of richer sentences could result from experimenting with transformer-based decoders [2]. The system would be more widely available with cloud-based deployment, enabling users to create captions without the need for local setup. Furthermore, multilingual text and speech output can increase accessibility and usability globally. In the future versions, real-time video transcription & interactive feedback mechanisms can be added to adjust & increase the performance of the solutions over time, using user interaction. Finally adding visual tools to show where the model is attending to in different areas of an image, will enhance transparency & build trust in AI-based Solutions. This will be especially important for applications that are focused on assisting people with disabilities.

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 15, Issue 2, February 2026

|DOI: 10.15680/IJIRSET.2026.1502087

REFERENCES

[1] S. Liu, L. Bai, Y. Hu, & H. Wang, "Image Captioning Based on Deep Neural Networks," MATEC Web of Conferences, vol. 232, 2018.
[2] F. M. Shah, M. Humaira, J. Jim, A. S. Ami, & S. Paul, "Bornon: Bengali Image Captioning with Transformer-based Deep learning approach," arXiv preprint arXiv:2109.05218, 2021.
[3] J. Aneja, A. Deshpande, & A. G. Schwing, "Convolutional Image Cap- tioning," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2018.
[4] Y. LeCun, L. Bottou, Y. Bengio, & P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, vol. 86, no. 11, pp. 2278- 2324, Nov. 1998.
[5] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Comput., vol. 9, no. 8, pp. 1735-1780, 1997.
[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2012, pp. 1097-1105.
[7] A. Farhadi, et al., "Every picture tells a story: Generating sentences from images," in Proc. Eur. Conf. Comput. Vis. (ECCV), 2010, pp. 15-29.
[8] G. Kulkarni, et al., "Babytalk: Understanding and generating simple image descriptions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 12, pp. 2891-2903, 2013.
[9] T.-Y. Lin, et al., "Microsoft COCO: Common objects in context," in
Proc. Eur. Conf. Comput. Vis. (ECCV), 2014, pp. 740-755.
[10] R. Krishna, et al., "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," Int. J. Comput. Vis., vol. 123, no. 1, pp. 32-73, 2017.
[11] H. Ayesha, et al., "Automatic medical image interpretation: state of the art and future directions," Pattern Recognit., vol. 114, 2021.
[12] Y. Pan, et al., "Automatic image captioning," in Proc. IEEE Int. Conf. Multimedia Expo., 2004.
[13] V. Ordonez, G. Kulkarni, and T. Berg, "Im2Text: Describing images using 1 million captioned photographs," in Proc. Adv. Neural Inf. Process. Syst. (NIPS), 2011, pp. 1143-1151.

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