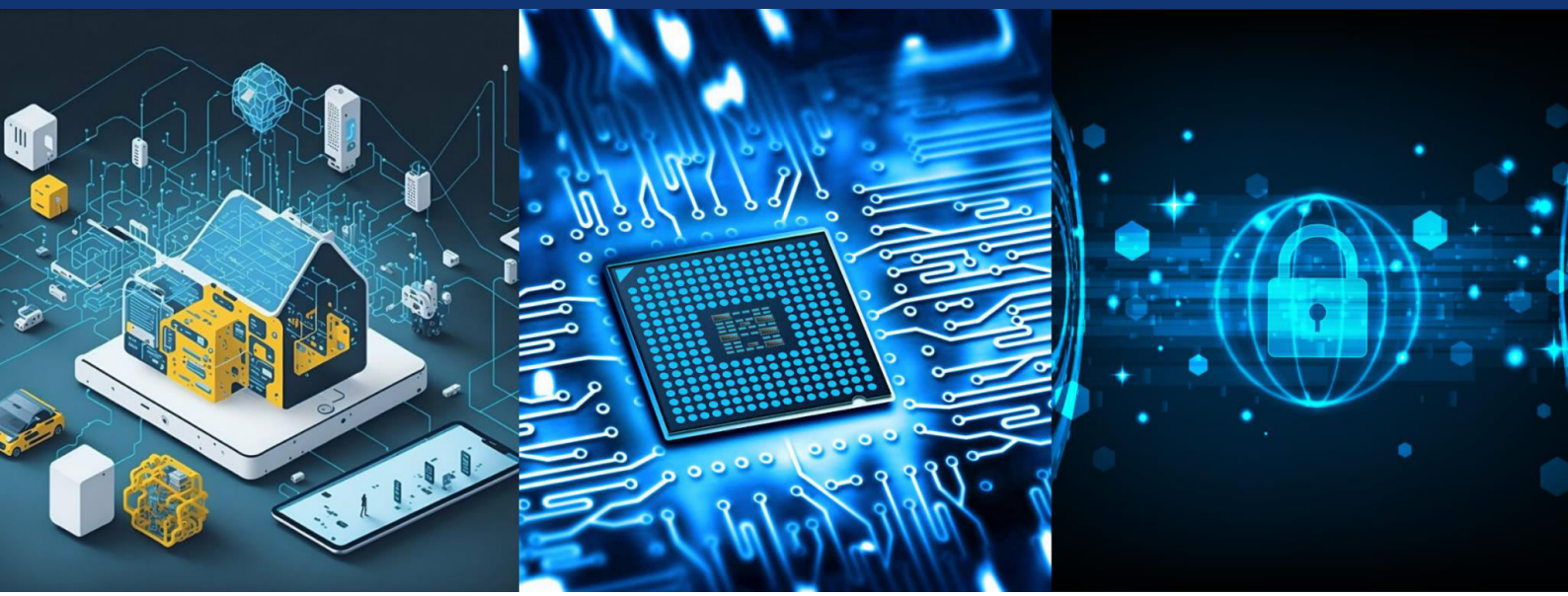
IJIRCCE

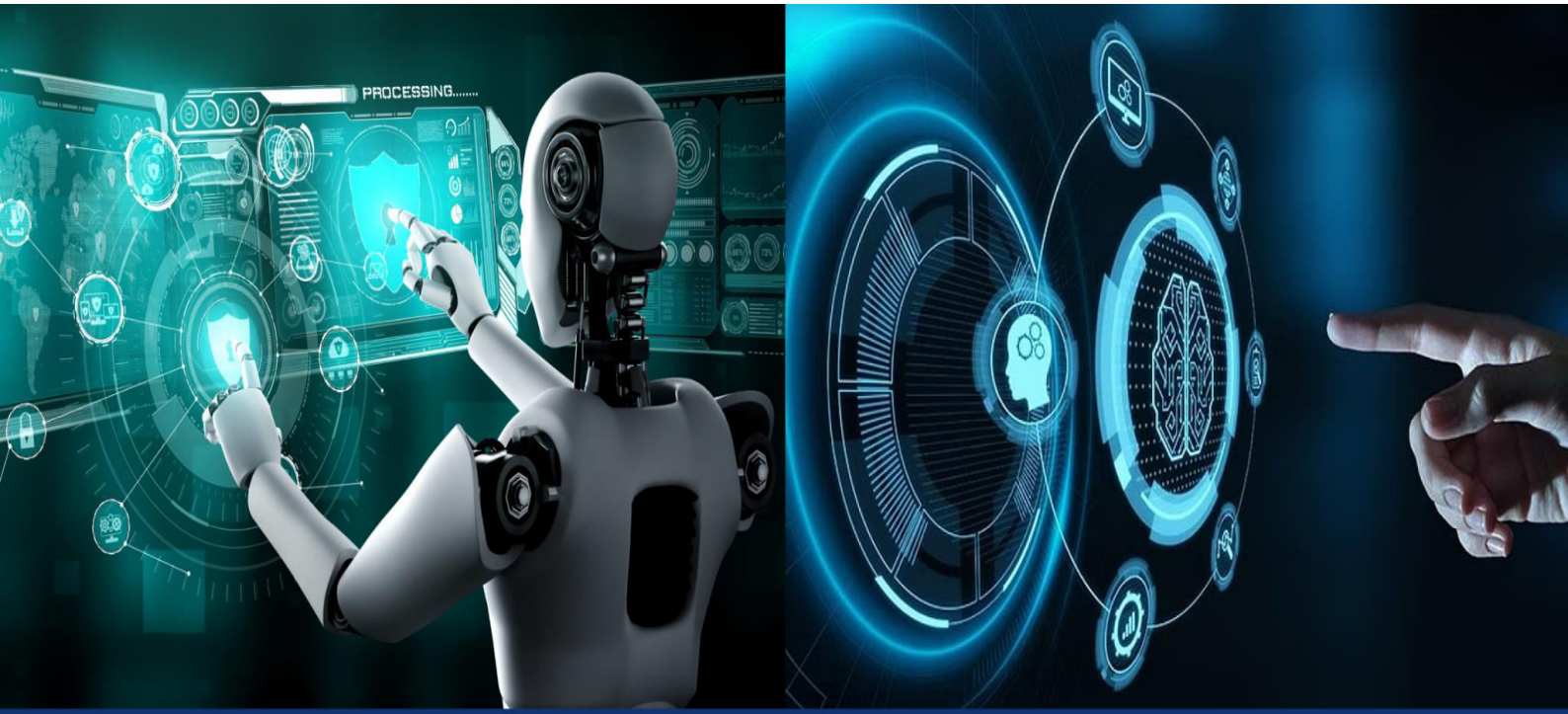
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

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



Impact Factor: 9.927

Volume 14, Issue 7, July 2026

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Deep Learning-Based Image Caption Generation Using CNN-LSTM with Beam Search and Greedy Search

[1]Chava Santosh, [2]JA Paulson

M.Tech Student, Department of CSE, Vara Prasad Reddy Institute of Technology, Palnadu Dist, AP, India

Associate Professor, Department of CSE, Vara Prasad Reddy Institute of Technology, Palnadu Dist, AP, India

ABSTRACT—Image captioning is one of the most important topics in the field of Artificial Intelligence (AI). That is, researchers and software developers have placed a great deal of attention on it. The way to create the deep learning method for understanding and describing images based on the situations is image captioning. This technology of this invention has versatile applications, such as in medical imaging, education or e-learning and that is why it is marked as a game changer. The sole function of an image caption generator is to use the images given to them to independently produce structured English sentences. Nevertheless, generating very accurate captions is still a very difficult process. The main idea of this paper is to develop a new approach to generating image captions that uses two optimization methods: beam search and greedy search. The system suggested is a program taking an image as input, and the output of the program is an English sentence that describes the picture content. The construction interconnects a pre-trained convolutional neural network (CNN) and VGG16 model to bring out the meaningful image characters and the introduction of long short-term memory(LSTM) network to generate the textual features. By employing these networks, the product is a powerful decoder able to effectively produce exact and contextually meaningful captions. Beside that, it is enabled to generate video captioning too by adding modules such as Convolution 3D (C3D) which allows the system to work through the frames of the images consecutively. The proposed framework mainly focuses on overcoming the innate problems of image captioning such as accuracy, ethical considerations, and responsible data usage. Persistent research together with the ethical use of the technology is required to maximize the beneficial aspects of this technology while managing its negative consequences.

KEYWORDS—Image Captioning, Deep Learning, Computer Vision, Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), Video Captioning

I.INTRODUCTION

Cutting-edge progress in Artificial Intelligence (AI) in the previous ten years has presented an enormous impact in diverse fields, including the development of image captioning systems as one of the most noticeable areas. Image captioning, a subcategory of computer vision and natural language processing (NLP), is about producing human-like understandable descriptions of visual content. The function FAI technologies to not only gauge the content of images but to construct coherent textual descriptions that bear similarities to human eyes, the one that would be interpreted from the visual scene; consequently, the machine needed to perform this task. The technology behind image captioning is the combination of deep learning models, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), with NLP techniques that come up with cognitive connections between the visual world and human language.

Image caption generation has extensive applications, ranging from the addition of social media content to the improvement of accessibility for the visually impaired. In the domains of healthcare, education, and entertainment, the technology may provide a tool for providing accurate, contextually appropriate captions, thus eliminating communication obstacles and leading to better comprehension.[1] For instance, the automatic captioning system can make a blind person understand a picture or video by presenting a descriptive text; in this way, they will be able to engage with digital media more easily. Moreover, the use of AI in e-learning will be advantageous as captions can be converted to different languages to cater to diverse needs in the communities of people who want to learn a new

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

language. Also, AI-generated captions can be applied to various techniques such as the analysis of medical imaging, e-learning, and online content creation, thus, by making the information and interactivity easily accessible.

The purpose of the study described in this paper is to work out an image captioning algorithm that can autonomously generate sensible and contextually correct captions in English based on the images entered. This model exploits the strength of deep learning techniques that interplay with pre-trained CNN architectures specifically with VGG16 and LSTM networks in a Long Short-Term Memory (LSTM) cluster. The CNN module is in charge of extracting the most important information in the input image, and the LSTM network decodes those features into meaningful sentences that reflect the content of the image. The reason for choosing VGG16 as the CNN model was its capability to extract features as seen from the fact that it has a large architecture with a deep level of understanding that can process complex patterns and exact details of the image.

Despite a substantial number of scientific papers attesting to the high level of efficacy of the CNN-LSTM tandem in the domain of image captioning, it is clear they have limitations. The selection of the CNN that is used to extract the features is one of the central problems that still resist. The models such as ResNet, Inception, as well as EfficientNet, have been the most frequently used but each has in respective fields qualities and flaws such as capturing finer details of the image, being computationally efficient, and being able to generalize to different image types. An experiment was performed to demonstrate that existing systems can be supplemented with the help of more advanced CNN architectures such as VGG16 in order to observe if they are capable of extracting important features more efficiently. VGG16, which is a bit older than the most recent models, is still one of the best choices because of its simple but highly effective design, which helps it to capture the essential visual features that are necessary for generating accurate captions.

The scope of the project is not just limited to image captioning, but also goes into video captioning which deals with the problems of generating describe the looks of the dynamic, time-dependent content. Video captioning, which involves recognition of the temporal dimension of video data, poses unique challenges, especially in the aspect of processing and integrating information across different frames. For this purpose, the present work introduces the use of Convolution 3D (C3D) modules which can process spatio-temporal data in videos. The new model with C3D captures motion patterns in the video to a better extent, thus helping the model to produce captions that reflect actions, series of events and so on more reliably. Hearing impaired people will benefit especially from this; it will be an aid to provide the text for events they are watching and thus, get easier access to the video content.

Though image captioning and video captioning have the potential of accessibility and usability, it should be noted that the process of development and deployment of such systems involves a number of complex, ethical and practical tasks. To steer clear of these two entities and guarantee the systems' operability, accuracy, and responsibility towards users, the challenges have to be taken care of. The real-time capability of the model to generate high-quality captions appears to be a major problem. These types of image captioning models, especially the ones that employ deep learning methods, need a huge amount of CPU time for processing images and generating captions efficiently. The balance between model performance and computational efficiency stands as a core point of ensuring the system runs fruitfully on different gadgets from high-end servers to mobile phones.

Together with the specialized difficulties ethical challenges also have to be considered. The construction of AI systems for creating captions brings up some vital questions about data sequestration, bias, and the liability of misinformation. Since the caption generation systems generally are grounded on large image datasets along with the corresponding captions, it's important to examine how these datasets are made and what kind of impulses are in them. Non-objective captions can either circulate misconceptions or display images in a deceiving way, with the result of inaccurate or dangerous content. Besides that, the composition also argues for further work on ethical AI with a focus on translucency, fairness, and responsibility in the training and deployment of image captioning systems. It's of utmost significance to produce systems that can minimize impulses and produce inclusive and different captions.

Among the main difficulties of espousing image captioning systems is their personalization following druggies' own wishes and inclinations. To this end, the ongoing design innovated a stoner-centric bracket approach which allows the druggies to train the model grounded on their own disciplines or requirements. By controlling the

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

machine to classify images into different orders like healthcare, education, and entertainment, the tool can induce captions that aren't only accurate but also are suitable for the sphere in question. This process, in turn, results in a further protean working of the image captioning model, thereby setting it piecemeal as a well- established instrument usable in a broad diapason of sectors, from social media disciplines to content creation tools to the health and education sector.

This paper aims to contribute to the advancement of image captioning technologies by developing a model that pushes the boundaries of what's presently possible. By combining advanced CNN infrastructures with LSTM networks, expanding into videotape captioning with C3D modules, and considering ethical and stoner- centric design considerations, this work aims to develop a more robust, accessible, and responsible image captioning system. In doing so, it seeks to contribute to the ongoing development of AI systems that can understand and induce language in a manner that nearly glasses mortal capabilities while considering the broader societal and ethical counter accusations of similar advancements.

II.LITERATURE REVIEW

It has made giant strides with deep learning to come to the field of image captioning, the field of generating really describing and contextually accurate captions for images[1][8]. It is one of those crucial tasks that will try to synergize computer vision and natural language processing in terms of producing captions, describing the content of an image in a mortal-like way.[1] Previously, the work on image captioning would have concentrated on rule-based and template-driven styles. However, with the advent of deep learning, the field has undergone a massive transformation. This literature check reviews the development of ways in image captioning, popular methodologies, and contemporary challenges in the field.

In the beginning, image captioning systems adopted hand-crafted features together with shallow learning models and methods as template based approaches and keyword matching for inducing captions.[2] However, these early systems struggled to produce rich, contextually accurate and grammatically correct descriptions.[3] The traditional styles depended on pre-defined templates or manually defined features that project the system to only be able to cope with a limited range of images in varied surrounds. These limitations generated the exploration of more advanced learning models in posterior times.[4] The earlier styles, although foundational, could not penetrate the depth needed for understanding complex visual data and the nuances needed in the content of images, resulting in captions that were all too often general or inapplicable.

Yet, it is the preface of the deep literacy models like Convolutional Neural Networks (CNNs) that revolutionized image captioning.[5] These are the models that reportedly excel at capturing the spatial scales and visual patterns within images.[6] Models such as AlexNet, VGG, and ResNet became one of the pivotal inventions that helped in the progress and development of any image captioning model further, such as the integration of attention mechanisms. Introduced by Xu et al., such attention mechanisms enabled the model to pay attention to separate regions of an image while being generated at every word in the caption.[7] This has added an extra layer of perfection to the model's ability to generate context-relevant captions. [8] Rather than treating the image as a single entity, attention mechanisms have allowed the model to pay more attention to the parts of the image that are more relevant to the word currently being generated, thereby improving the general quality of captions. This development has greatly enhanced the ability of the model to generate richer descriptions that align better with human perception of visual content.

The emergence of Motor Models is another touchstone in the history of image captioning. Such an example is the Vision Transformer(ViT), into which one may insert a self-attention mechanism that captures global dependencies within the image.[9] Motor models, unlike a traditional localized CNN point detector, may model extensive dependence and connection between portions of a given image.[10] Moreover, the aforementioned conversion from CNN to Motor- based architectures resulted in significant improvements in the performances of captioning with respect to landing of situational associations and understanding fine-grain details in the image.[11]In addition, Mills introduced the provision of ressembler processing, thus making these models more amenable to very heavy datasets and scarves.

Recent developments in image captioning now present visitors with thick captioning, a style of captioning that

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

creates more than one caption for an area in the image, thus offering a wider and fuller description of what the image is all about.[12] It is indeed quite useful for such applications which have an object in understanding several different objects or scenes inside a single image, thereby allowing for a wider-scoped comprehension of the complete visual content.[13] Also, transfer learning by coupling pre-trained models such as GPT-3 and BERT has become widely adopted in image caption generation tasks. These are mostly NLP-intended models, adapted to grasp and generate textual-based on the visual features extracted by CNNs. This strengthening of visual and verbal models has resulted in a huge amplification in the quality of generated captions; they are becoming more contextually accurate and semantically.

Another important aspect has been multimodal models like CLIP (Contrastive Language-Image Pre-training), which appear to combine textual and visual data in a much more tangled manner. CLIP models are trained over huge amounts of images and text descriptions, enabling them to infer more semantic captions. Due to joint learning of both image and text data, these models can relate much better between the two modalities and thus improve the title accuracy and contextual applicability.

Typically, an image captioning model has an automatic evaluation similar as that of BLEU, METEOR, ROUGE, and CIDEr.[6] These criterions assess the quality of captions with reference to human-generated reference captions. These metrics focus on reconciling n-grams and similar semantics and give a quantitative check on the quality of captions. [3]Automated criteria, however, are not exhaustive, and it is the human evaluation that is most appropriate for qualitative aspects such as coherence, applicability, and vagueness.

Here, the human evaluator can appreciate more the nuances of the context and meaning, which are essential for producing a more human-like caption.

Although substantial advances have been witnessed in image captioning, many issues remain in the field. It is, for instance, a big problem to cope with the ambiguity that arises from the different interpretations that can be made out of an image to generate many appropriate captions. Another continuing problem is to be able to generate different captions which are contextually appropriate, yet not repetitive.[14] More advanced methods such as scene graph generation and hierarchical captioning may provide better structured representations of content in the image to mitigate some or all of these issues.

In summary,the field of captioning images has since undergone tremendous changes ranging from the use of deep learning models, attention mechanisms, transformers, and multimodal models. As with so much improvement, the generated captions have also improved in terms of their quality, becoming more contextually and linguistically apt and diverse. There are still some areas for research, such as ambiguous captions; captions of diverseness; and the inclusion concept in contexts that can be explored. As the field moves forward, it would show even more in this regard, specifically by determining such improvements that have to be feasible for generating significant captions by models for a wider range of real-world applications.

III. PROBLEM STATEMENT

The recent ages of digitization have demands towards proper understanding tools in interpreting visual data. It would be passing descriptive captions generated for images which would do an excellent job in closing the gap between visual content and textual understanding. Machines would then learn to comprehend and convey visual information in a manner that resonates with human interpretation. Benefits particularly accrue in accessibility in which a visually impaired person perceives as understanding through a textual description. There are also other domains like medical imaging, e-learning, and automated content generation.

These days, advancements in artificial intelligence are not able to help with the generation of image captions, which will not be found tagged with any specific names. Many challenges arise while interpreting visual items, establishing their relation, and finally writing grammatically-correct sentences that represent the content of the image. Because of the above-mentioned reasons, approaches originating from traditional methodologies prove to be lesser in both accuracy and contextual relevance. Thus, there is always the need to look for better and boisterous methods that can do the job.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

We're developing an image caption generator that would have as its core the application of deep learning methods to integrate Convolutional Neural Networks (CNNs). CNNs work very closely with Recurrent Neural Networks (RNNs). Towards the end, the project will develop an annotation-generating system that will comprise a pre-trained convolutional neural network, the VGG16 architecture, which will extract high-level features from images, understanding the characteristics needed. These features will then be coupled with layers of Long Short-Term Memory-an RNN-specialized type-with reference to processing sequential textual data to generate meaningful captions.

It uses optimization techniques like beam search and greedy search in the project in improving the generated captions' accuracy and quality. Further, this model includes the integration of VGG16 for extraction of features and LSTM for sequential generation through captions that are descriptive and contextually relevant in their generation.

The broader scope of this technology includes video captioning, where modules like Convolution 3D (C3D) can be integrated for handling the temporal component of the data. Image captioning has lots of difficulties in terms of research and exploration regarding its efficiency, bias mitigation, and ethical use of data resources. This project strives to take a step forward in developing image captioning technology which can revolutionize application areas.

IV. PROPOSED SYSTEM

An automatic image caption generation system is proposed for the generation of automatic descriptive and contextual captions for images with advanced models by deep learning techniques. The proposed system generates captions that are precise and coherent with the images, using feature extraction, sequence modeling, and optimization methods. The following describes the various components and working behind the proposed system.

1. Image Input and Preprocessing

- Image Upload: The first step of the system involves a user uploading an image; the system is capable of accepting different image formats into the image upload module, such as JPEG, PNG, etc., for it to be processed.

- Preprocessing: After an image has been uploaded, preprocessing makes sure the image is standard to the deep learning model. It adjusts the image to a fixed standard, which is mostly according to the input requirement of the Convolutional Neural Network (CNN) model, which may be 224 by 224 pixels for 'VGG16'. Then normalization across images is performed to have its pixels normalized within a fixed range, e.g., between 0 and 1. To make the model robust against various conditions of the images, data augmentations like random rotations, flips, and scaling can be applied. Hence, this would contribute to better performance of the model in the real scenario.

2. Feature Extraction using CNNs

- CNN architecture: The system is using pre-trained CNNs, like VGG16, ResNet, Inception, or EfficientNet, to extract high-level visual features from an image. These models are used for image popular word patterns because they have been studied and found to capture very complex patterns in the input images and perform well as well during the image recognition tasks. Each of the various CNN models is equipped with their very own abilities; for instance, ResNet is known for its ancestry learning, which helps train deeper networks, while EfficientNet optimizes both accuracy and efficiency in terms of computation.

- Feature Maps: The image from the CNN passes through the neural network in such a way that lower level briefly learned features include edges or textures and the higher level learned more complicated objects and their spatial hierarchies. The output of a CNN is a collection of feature maps, which are typically high-dimensional representations of the image that encode the content of that image. These feature maps can then be used to generate the caption.

3. [12]Sequence Modeling with LSTM. Long Short-Term Memory (LSTM) Network: The derived feature maps are fed to an LSTM network for sequential data modeling, which happens to be optimal for language generation tasks. The LSTM network takes the visual features from the CNN and starts generating a word sequence, leading to a caption.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Embedding Layer: An embedding layer is introduced within the LSTM to convert words into dense vectors representing the semantic relationships between them so that the system understands context better. This is because the embedding layer also enhances the system's capability to generate captions that are syntactically correct and semantically meaningful.

Attention Mechanism: For further improvement in caption accuracy, an attention mechanism is integrated into the LSTM network. When generating each word of the caption, it enables the model to focus on different parts of the image. For example, while producing the word "cat," the attention mechanism enables the model to look at that portion of the image where the cat resides, which produces a relevant caption with better precision.

Caption Optimization

4. Beam Search: The process of decoding an image through the LSTM entails a Beam Search, which is what the system employs in capturing quality captions. This is reduced to the fact that when decoding, Beam Search considers several word sequences at a time and ranks words on their likelihood to come up with the best caption. It spans the space of many sequences without burdening computation since it operates like most sequences without an increase in load.

- Greedy Search (Alternative): The system is able to use the Greedy Search that chooses the most probable word at each step, without taking into consideration the alternative-way sequences. This is the reason for this alternative: Fast and resource efficient, but ends up producing captions that may not be as optimal compared to Beam search.

5. Customization of the User

- User Input: In the same regard, a user may indicate their preferences for a particular tone, formality, or specific keywords; for example, to issue captions that are formal or that have specific terms.
- Caption Refinement: In this respect, the system changes the captions generated with respect to user preferences. Thus, the user may get captions that relate not just accurately to the image but also match with their stylistic and contextual paradigms.

6. User Review and Feedback

- Caption Review: Review the caption generated by the system and have an option for the user to provide feedback. In this iterative process, the user requests alteration of the caption if it does not match their expectations.
- Learning from Feedback: The user feedback is stored and, henceforth, used to improve the output. This keeps on improving the overall performance of the system, adapting new requirements based on user preferences and creating personalized captions for them later.

7. Export and Integration

- Social Media Integration: Once the caption is approved by the user, the system automatically exports both the image and its caption to the user's selected social media platforms such as Facebook, Instagram, or Twitter. This seamless integration takes away the cumbersome process of captioning for sharing contents on the sites.
- API support: The system provides API access to developers in order to facilitate the system integration with any other application or content management service, thereby opening a greater possibility to embed the captioning system in many more platforms or services.

8. System Training and Evaluation

- Training Data: The system is trained on very large datasets, each containing images with several captions. Such datasets can make the model learn many-to-many image-caption pairings, thus being able to describe numerous images.
- Evaluation Metrics: [6]The performances of the system on the output side are evaluated by measuring the standard such as BLEU that prescribes the n-gram overlap of generated captions with those of references. All of these metrics measure the accuracy of the generated captions from a linguistic and contextual perspective.

Hence, the proposed system for image caption generation houses advanced deep learning models such as CNNs, LSTMs, and attention mechanisms to deliver a powerful set of highly accurate and contextually relevant

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

captions. Through customization options, user feedback, and continuous learning, the captioning system will serve multiple applications of usage in a personalized yet efficient manner

V. METHODOLOGY

This report is supported by a rich methodology spread across various parts of the report. This is a summary of the structure of the methodology based on the design and testing of the project:

1. System Design and Software Requirements Image Input Handling:

-The system must accept standard image formats, such as JPEG, PNG, BMP, and so on, for transfer of any images and be easily compatible.

- Preprocessing of images must comprise resizing, normalizing, and augment-ing to have uniform input quality meant for better extraction of features.

Caption Generation:

- Uses a deep learning model built with CNN for feature extraction and an RNN or [2] Transformer architecture for generating captions sequentially from this feature extraction.
- Capable of producing captions for images in real-time or near real-time, significantly reducing the delay imposed on users.

User Interface:

- This will provide a simple and easy-to-navigate interface capable of uploading images and receiving captions.
- It also has space for user-feedbacks regarding captions, ensuring the model gathers complete data for improving through iterations.

Model Training and Fine Tuning:

- Useful tools for training the model on new datasets for its requirements and specific competencies that need to be met in the future.
- Facilitates such database management for easy update and enhancement of the model.

Integration:

- APIs provide the endpoints to integrate image captions with any external applications and services.
- Compatible with all the platforms for easy usage and deployment.

Non-Functional Requirements

1. Security:
   - It's serious about ensuring total confidentiality, integrity, and security of handling and will thus also protect the user's privacy.
2. Performance:
   - With the response time to generate captions remaining low, it will give the user an experience based on efficient interaction.
3. Reliability:
   - Minimal downtime-they-designed-for this to ensure that when it runs, it runs in most environments all through.
4. Usability:
   - The simple interface user will be able to navigate through using the application without having to go through lengthy training periods.
5. Scalability:
   - This implies that it can adjust to increasingly large dataset sizes, functionalities, and users without the system's performance levels being affected.
6. Maintainability:
   - It should be documented thoroughly by the latter to easily allow upgrades, debugging, and changes in future, enabling the use of applications for many years.

Design Software

Unified Modeling Language (UML) illustrates Architectural and Functional Diagrams that represent the system architecture and several functionalities.

Structural Diagrams: Class and component diagrams represent the model's components and their connections.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Behavioral Diagrams:
Sequence and activity show the workflows, explaining which data is flowing through the systems and how interactions are accomplished.
Implementation Features:
State diagrams have been added to depict changes over time in the state of the system during any kind of user interaction.
Workflow:
The work flow depicts the generation of captions from user images to social media. It comprises feature extraction using CNN, caption generation using LSTM, optimization using Beam or Greedy Search, and user customization. This process ends with user feedback cycles for improvements and eventually sharing successful results on social media.

2. Encoding and System Deployment

The inputs are then processed into photonic feature representations through a pre-trained Convolutional Neural Network (CNN),[13][15] such as VGG16 as a feature extraction model from raw pixel data, and then fed into a Recurrent Neural Network (RNN) or Transformer model like LSTM or attention mechanism, to generate coherent and contextually relevant captions. During preparation for training, preprocessing operations include image resizing, normalization of pixel values, and tokenizing the associated text captions.

During this phase, the trained model is integrated into a deployable accessible application for users. It converts the model into a format that matches the deployment platforms with appropriate libraries such as TensorFlow Lite, ONNX, or Keras H5. An application like Streamlit is used to create web service applications for the interaction between users and prediction results.

Once the user inputs an image, the system will run the image-through the pre-trained model, extract certain features, and decode these features to generate a caption in the deployed system. To allow for real-time inference, optimization is done through model quantization and load balancing across servers. Given that this scenario will represent a number of simultaneous requests flowing through the real-world servers, the system should maintain high uptime, quick response times, and efficient usage of resources.

Thus, encoding refers to feature extraction and sequence modeling, while deployment deals with usability, i.e., scalable web interfaces through which users can interact seamlessly with the image caption generation system.

3. Testing

Testing Techniques
White Box Testing: Test internal workings of a system; test for logical paths and code structure
Black Box Testing: Functionality testing from the point of view of a user. Test that the outputs produced match the expectations.
Gray Box Testing: It tests the functionality along with the limited internal structures and can be of much use while testing for integration.
Types of Testing:
Unit Testing:
Test every module or function independently to test every individual unit.
Integration Testing:
This testing ensures the different modules are integrated, which further helps to get smooth output from all components of the system.
Regression Testing:
Ensure the newly developed functionalities do not harm the already built functionalities.
Performance Testing:
Check response time, processing speed, and memory usage. Stress Testing:
It tests the model for high load conditions as well as its robustness at peak demands.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

Security Testing:
This tests for user data protection as well as guarding against unwanted access into the model.
Usability Testing:
It checks how easily accessible it is for a user in terms of navigating and interacting.

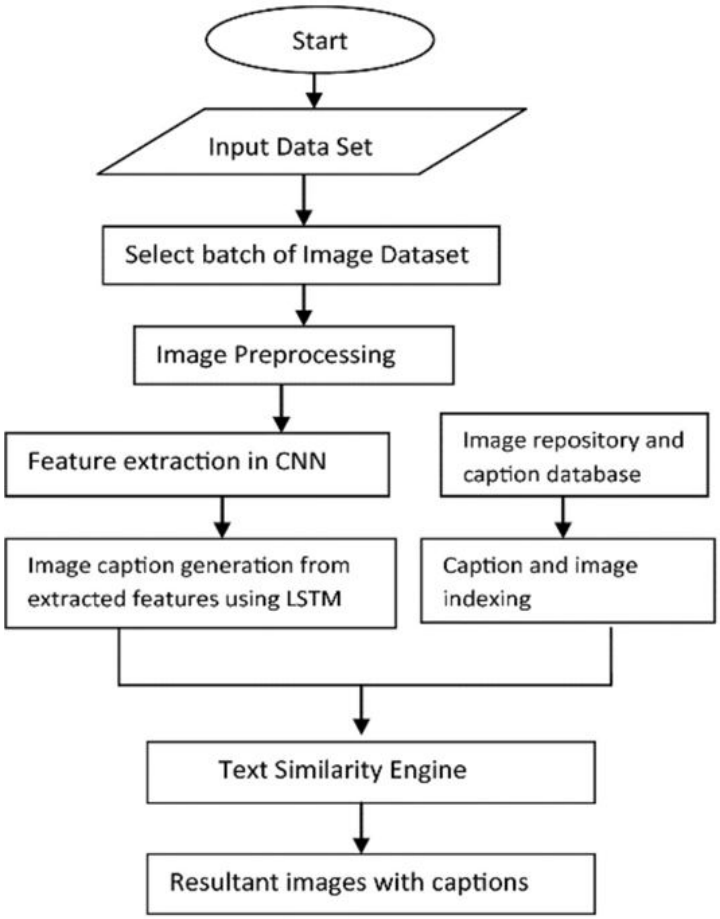


Fig. 1. Work Flow of the Model

VI. RESULTS

In contrast to CNN-RNN models like VGG16 combined with LSTMs, which show continued success for image captioning, the majority of machine learning tasks tend to be dominated by transformer-based models. However, this proves to be a much better balanced approach for visual feature extraction with sequential capabilities in captioning. It is, therefore, an appropriate approach for use in resource-limited environments. A series of captions is therefore generated from the model and ranked based on correctness. They depict the content of the image either in terms of objects, actions, or emotions. Their relevance is also crucial as they define what things matter in an image. And of course, there is room for diversity in the captions, which can be manipulated based on settings, such as temperature, affecting creativity but at the expense of accuracy with higher temperatures.

These models are very problematic at supplementing visual information with temporal sequences. By the CNN component, features from the image are extracted, which are subsequently fed into the RNN in order for the caption

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

to be produced word by word. Due to the decreased computational power, these types of models outperform transformer-based models and yield good performance with less data. Additionally, they may face problems because of the inherent feature of RNNs, that is, they would have problems in long-range dependencies, which many of such issues may be solved when one uses LSTM networks.

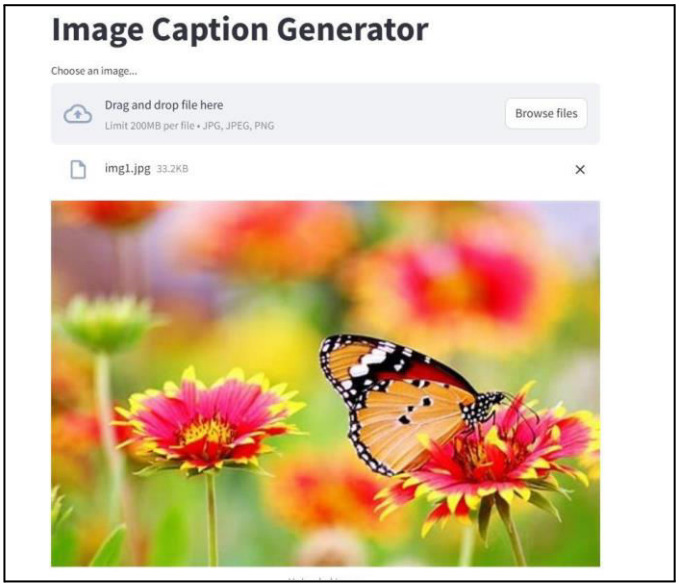


Fig.2.Image Input

The effectiveness of image captioning models can be evaluated through many measures. One among them is the BLEU score, which matches generated captions against reference captions and finds the overlap of n-grams. Other measures like ROUGE, METEOR, and CIDEr measure the quality of the captions as well as their similarity to human annotations. Model performance is dependent on several factors, including the variety in the training dataset, the network architecture, and how training is done. Overfitting or having a model that generalizes very well on training data but fails miserably when new images are introduced is a common challenge that needs proper validation to ensure good performance in future use of the models..

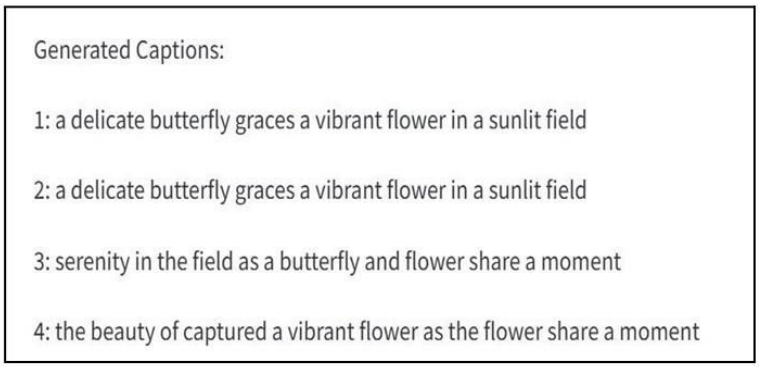


Fig. 3. Output captions in the selected genre

They will be syntactically correct and semantically meaningful captions according to the effectiveness of the image captioning models, and they will have better scores of evaluation measures.

© 2026 IJIRCCE | Volume 14, Issue 7, July 2026 DOI: 10.15680/IJIRCCE.2026.1407006

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

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

VII. CONCLUSION

This paper demonstrated that combining CNN-based feature extraction with LSTM-based sequence generation provides an effective framework for automatic image caption generation. The comparative evaluation showed that both Greedy Search and Beam Search produce coherent captions, with Beam Search generally achieving higher-quality and more contextually accurate descriptions. The proposed approach enhances semantic understanding by effectively mapping visual features to natural language. These findings highlight the potential of deep learning models for applications such as assistive technologies, image retrieval, and multimedia content analysis. Future work may focus on integrating transformer-based architectures and larger multimodal datasets to further improve caption accuracy and contextual relevance.

REFERENCES

[1] Roshan Adhithya SS, Priyadharshini M, Lekshmi Kalinathan (2023). Image Caption Generation for Blind Users Of Social Media Websites
[2] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, Rita Cucchiara (2020). Meshed-Memory Transformer for Image Captioning by Cornia et al.
[3] Shuai Li, Zhihao Tao, Kai Li, Yun Fu (2019). Visual to Text: Survey of Image and Video Captioning
[4] Grishma Sharma, Priyanka Kalena, Nishi Malde, Aromal Nair, Saurabh Parkar (2019). Visual Image Caption Generator Using Deep Learning
[5] Pascal Dognin, et al (2019). Adversarial Semantic Alignment for Improved Image Captions.
[6] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang(2018) . Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.
[7] Jiasen Lu, Caiming Xiong, Devi Parikh, Richard Socher (2017). Knowing When to Look: Adaptive Attention via a Visual Sentinel for Image Captioning.
[8] Justin Johnson, Andrej Karpathy, Li Fei-Fei (2016).DenseCap: Fully Convolutional Localization Networks for Dense Captioning.
[9] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan (2015). Show and Tell: A Neural Image Caption Generator.
[10] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio (2016). Image Captioning with Semantic Attention.
[11] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio (2016). Neural Image Caption Generation with Visual Attention.
[12] Ali Ashraf Mohamed (2020). Image Caption using CNN and LSTM.
[13] Palak Kabra, Mihir Gharat, Dhiraj Jha (2022). Image Caption Generator Using Deep Learning.
[14] Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur (2021). Image Caption Generator.
[15] Sreeja Pinnapureddy, Abhishek Raparthy, Reethusree Suthari, A. BhanuPrasad, (2023). Enhanced Image Caption Generator using Deep Learning.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details