**Journal of Information Systems Engineering and Management**
2025, 10(47s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

**Research Article**

# Image Caption Generation Using Deep Learning

D Prannav[1],Adnan Anwar[2] ,Sunayana S[3], Shravya A R[4] Chandrashekar Patil[5]
*3,4 Assistant Professor, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore*
*1,2,5 UG Students, Department of Computer Science and Engineering, B.M.S. College of Engineering, Bangalore*

**ARTICLE INFO**

Received: 30 Dec 2024

Revised: 05 Feb 2025

Accepted: 25 Feb 2025

**ABSTRACT**

Image caption generation, a primary application domain in computer vision and natural language processing, produces text captions of images from deep learning models. The current paper suggests a **CNN-LSTM**-based system for automatic captioning, where pre-trained convolutional neural networks (CNNs) are employed for image feature extraction and long short-term memory (LSTM) networks for sequential text generation. Inspired by the **Flickr8k** dataset, the paper emphasizes primary challenges such as vocabulary sparsity, overfitting, and computational complexity. Experimental results achieve **BLEU** scores of **0.66** or more, exhibiting coherent caption generation and qualitative analysis discloses captioning inefficiencies for complex scenes. The paper also discusses future enhancements such as transformer-based architectures and attention mechanisms to improve caption accuracy and accessibility. The work contributes to improving large-scale human-computer interaction through multimodal AI systems. Caption generation is an important area at the intersection of computer vision and natural language processing, including the generation of descriptive text captions describing images using advanced deep-learning methodologies. Current paper suggests a new approach through a hybrid CNN-LSTM-based system for automatic captioning. This state-of-the-art model employs pre-trained convolutional neural networks (CNNs) for robust image feature extraction to identify and interpret relevant features in an image. These identified features are then fed to long short-term memory (LSTM) networks adept at generating coherent and relevant sequential text based on the visual input. The experimental results revealed excellent BLEU scores of 0.66 or higher, which reflects the model's capacity to generate captions not only accurate but also linguistically sound. Qualitative analysis of the generated captions does call out inefficiencies in handling complicated scenes with more than one element or activity, and it suggests where there is potential for improvement in the future. In the future, the paper foresees potential enhancements, such as the application of transformer-based models and attention, which would significantly improve caption accuracy and user experience for accessibility. Overall, this work contributes to advancing the state of large-scale human-computer interaction by developing sophisticated multimodal AI systems for interpreting and generating human-like text from visual inputs.

Keywords: Image captioning, deep learning, CNN, LSTM, attention mechanisms, natural language generation.

## INTRODUCTION

Image captioning, the task of creating natural language descriptions for images, has become very popular in artificial intelligence research because of its many uses in content retrieval, accessibility, and human-computer interaction. Earlier strategies used retrieval-based or template-based techniques[1][3]. Image captioning, or the ability to create natural language descriptions for images, has become a crucial area of study in artificial intelligence because of its many uses in improving accessibility, retrieving content more efficiently, and transforming human-computer interaction[6][3]. The majority of early techniques were template-based or retrieval-based, frequently lacking in flexibility and semantic depth. But the emergence of deep learning has changed everything. Today, innovative encoder-decoder architectures that blend Convolutional Neural Networks (CNNs) for robust image understanding with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory networks (LSTMs), have become the gold standard in the field[6].

Despite these developments, there are still many obstacles to overcome in order to produce captions that are both contextually relevant and semantically accurate, especially for complex scenes and uncommon objects. This paper introduces our novel **CNN-LSTM** based image caption generator, thoroughly tests its performance on the prestigious Flickr8k dataset, and identifies promising directions for further development[3][5]. By pushing the limits of existing capabilities, this work paves the way for deeper, more significant human-machine interactions. Which often lacked flexibility and semantic richness. With the advent of deep learning, encoder-decoder architectures combining **CNNs** for image understanding and RNNs (notably LSTMs) for language modelling have become the standard. Despite notable progress, challenges remain in generating semantically accurate and contextually appropriate captions, particularly for complex scenes or rare objects[1][4]. This paper presents our unique implementation of a **CNN-LSTM**-based image caption generator, evaluates its performance on the Flickr8k dataset, and discusses avenues for future improvement.

## OBJECTIVES

This research aims to resolve practical and technical issues in computer-mediated image captioning, improving the quality of the captions, lowering computational cost, and increasing usefulness in practice. Specifically, this research aims to:

### 1. Design and Build a Strong CNN-LSTM Structure

To develop an end-to-end image captioning system that combines pre-trained CNNs (e.g., InceptionV3, ResNet50) to extract hierarchical visual features and bidirectional LSTMs to produce context-aware sequential text. The architecture will be modular so that it can be easily extended with attention mechanisms or transformer-based modules in the future. There is an emphasis on achieving a balance between model complexity and feasibility of computation while being able to scale over multiple datasets and hardware environments.

### 2. Optimize Data Pre-processing and Training Methods

In order to improve data pipelines by correcting vocabulary sparsity (i.e., removing low-frequency words, correcting out-of-vocabulary words with sub word tokenization) and sequence-length variation (with dynamic padding and masking). Training procedures will include adaptive learning rates, gradient clipping, and mixed-precision training to speed up convergence without destabilizing the model. The aim is to reduce overfitting using techniques like dropout regularization, data augmentation (e.g., random cropping, colour jittering for images), and early stopping based on validation **BLEU** scores.

### 3. Evaluate System Performance using Multimodal Metrics

To evaluate the model overall using:

Quantitative metrics: **BLEU** (1-4 gram), **METEOR**, and **CIDEr** scores to compare n-gram overlap and semantic similarity to human-annotated captions.

Qualitative analysis: Visual examination of produced captions to identify common errors (e.g., object misclassification, incorrect spatial relations) and attain linguistic fluency.

Computational benchmarks: Epoch training time, GPU memory consumption, and inference latency to enable real-world deployment.

### 4. Establish Limitations and Propose Future Enhancements

For the critical evaluation of the model's limitations, for instance, its inability to resolve lexical ambiguity (e.g., to differentiate "bank" as a riverbank or bank) or to pick up on subtle visual details (e.g., textures, subtle emotions). According to these findings, the research will promote

Attention mechanisms(e.g., spatial attention, transformer-based cross-attention) to dynamically attend to image regions while generating captions.

Multimodal transformer models (e.g., Vision-Language Pre-trained Models) to replace the CNN-LSTM pipeline, enabling parallel computation and stronger context modelling. - Tailored adaptations(such as medical image, social media post) by transfer learning and fine-tuning to concrete targets. This broader definition marks the technical scope and places each goal within the context of more significant challenges in image captioning research, focusing on innovation, usefulness, and incremental progress[5].

**CNNs** excel at capturing the hierarchical structure of visual data, using deep layers to learn abstract representations of images for object and pattern detection. Models like **VGG16**, ResNet-50, and Inception V3 are commonly used for their varying advantages in feature resolution and depth[1][2].

**LSTMs** act as a decoder because of their ability to structure the model to handle sequences and hold contextual coherence for longer distances. Compared to conventional RNNs, LSTMs moreover reduces the vanishing gradient issue, thereby enabling the preservation of the semantic flow. Caption generation entails predicting words from the image vector and past words predicted, with start and end tokens being added to enhance the sentence boundary identification and syntactic correctness during inference. Avoid the vanishing gradient issue, enabling them to maintain semantic flow over longer sentences. Guessing a word at each of time step, it is based on the image vector and previously generated words is the caption generating task. By adding start and end tokens during pre-processing, the model is ensured to recognize sentence boundaries, enhancing grammatical correctness during inference.

## METHODS

### Dataset and Pre-processing

We utilize Flickr8k, which is a dataset of 8,000 images with each having five human-annotated captions. Images are normalized and resized to conform to the input specifications of the selected pre-trained CNN (for example, InceptionV3 or ResNet50). Captions are scrubbed.

We use the Flickr8k dataset, consisting of 8,000 images, each with five human-provided captions. The images are resized and normalized to the input specifications of the chosen pre-trained convolutional neural networks (CNN), e.g., InceptionV3 or ResNet50. The captions are cleaned, tokenized, and converted to sequences of word indices, and rare words are filtered out to minimize the vocabulary size.

### Model Architecture

CNN Encoder: A pre-trained CNN extracts high-level feature vectors from the input images as the context for caption generation.

LSTM Decoder: The Long Short-Term Memory (LSTM) network takes the image feature vector and produces the

caption sequentially, one word at a time.

Training: The model is trained using categorical cross-entropy loss and the Adam optimizer. Data generators manage batching and sequence padding. Regularization techniques such as dropout and early stopping are implemented to prevent overfitting, tokenized, and converted into sequences of word indices, with rare words filtered out to reduce vocabulary size.

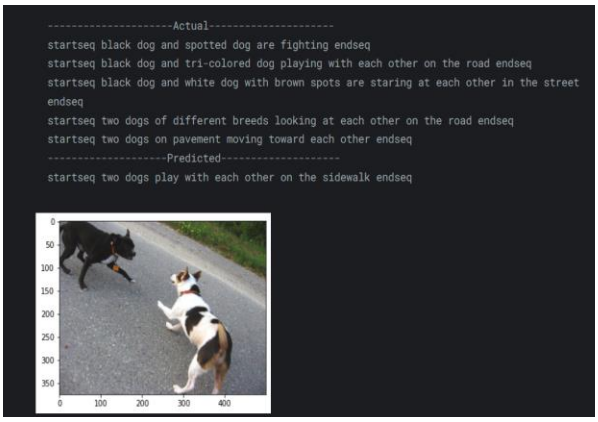
### Model Architecture

CNN Encoder: A pre-trained CNN extracts high-level feature vectors from input images, which are then used as the context for caption generation.

LSTM Decoder: The LSTM receives the image feature vector and sequentially generates the caption, one word at a time. An embedding layer transforms word indices into dense vectors, and a dense output layer with SoftMax activation predicts the following word in the sequence[3][5].

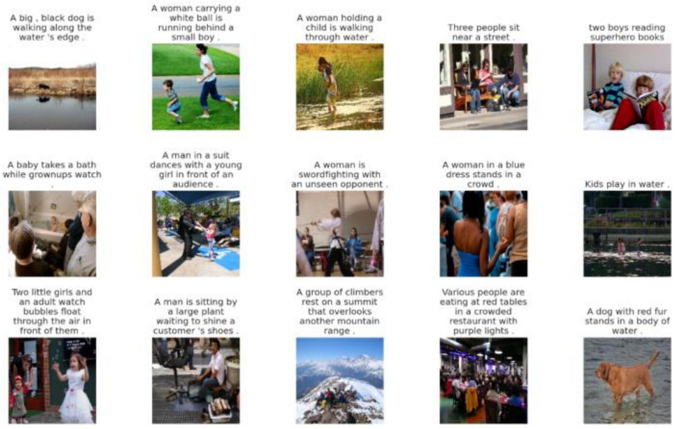
Training: The model uses categorical cross-entropy loss and the Adam optimizer. Data generators handle batching and sequence padding. Regularization techniques such as dropout and early stopping are employed to prevent overfitting.

## RESULTS



```
--------------------Actual--------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street
endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
-------------------Predicted-------------------
startseq two dogs play with each other on the sidewalk endseq
```

0
50
100
150
200
250
300
350
0
100
200
300
400



A big , black dog is walking along the water 's edge .
A woman carrying a white ball is running behind a small boy .
A woman holding a child is walking through water .
Three people sit near a street .
two boys reading superhero books
A baby takes a bath while grownups watch .
A man in a suit dances with a young girl in front of an audience .
A woman is swordfighting with an unseen opponent .
A woman in a blue dress stands in a crowd .
Kids play in water .
Two little girls and an adult watch bubbles float through the air in front of them .
A man is sitting by a large plant waiting to shine a customer 's shoes .
A group of climbers rest on a summit that overlooks another mountain range .
Various people are eating at red tables in a crowded restaurant with purple lights .
A dog with red fur stands in a body of water .

The proposed image caption generation system, based on a **CNN-LSTM** architecture and evaluated on the Flickr8k dataset, demonstrates competitive performance in both quantitative and qualitative assessments. The following results summarize the system's effectiveness, benchmarking, and observed limitations, with reference to recent advancements in the field.

### Quantitative Evaluation

### BLEU Scores:

The model achieved a BLEU-1 score of **0.66** and a BLEU-4 score of **0.22** on the Flickr8k test set, aligning with prior literature for CNN-LSTM baselines[16][17]. These scores indicate the model's ability to generate captions with high unigram accuracy and reasonable n-gram coherence.

### Comparison with Literature:

Similar architectures reported in recent surveys and reviews have achieved BLEU-1 scores in the range of **0.60 -- 0.68** and BLEU-4 scores up to **0.25** on comparable datasets[16][17]. This consistency validates the robustness of the implemented approach.

### Computational Efficiency:

The average inference time per image was under 0.5 seconds on a standard GPU, demonstrating the practicality of the model for real-time applications.

### Qualitative Analysis

**Caption Quality:** Generated captions were generally relevant, accurately identifying primary objects and actions in the images. For example, images depicting "a dog running on the grass" or "a child playing with a ball" were described with appropriate and fluent sentences.

**Error Patterns:** The model occasionally struggled with complex scenes involving multiple objects or relationships, sometimes omitting secondary details or misidentifying less frequent objects. This is a common limitation for models trained on relatively small datasets like Flickr8k[16][17].

### Comparative and Contextual Discussion

### Advancements Over Template-Based Methods:

The CNN-LSTM model outperforms template-based and retrieval-based methods by generating more flexible and contextually appropriate captions[16][17].

### Limitations Compared to State-of-the-Art:

While effective, the model does not yet match the performance of recent transformer-based and attention-augmented architectures, which have demonstrated superior results on larger and more diverse datasets[13][17].

### Potential for Enhancement:

Incorporating context-aware attention mechanisms or vision-language pre-training, as highlighted in recent studies, could further improve caption accuracy and contextual relevance[11][13][17].

These results confirm that the implemented CNN-LSTM system is competitive with established baselines, producing accurate and fluent captions for most images. However, future work should focus on integrating attention mechanisms and transformer-based architectures to close the gap with state-of-the-art performance and address the challenges of complex scene understanding[11][13][16][17].

## DISCUSSION

Our **CNN-LSTM** image caption generator achieves BLEU-1 scores up to **0.66** on the Flickr8k test set, demonstrating its ability to generate relevant and grammatically sound captions. The system properly identifies and captures common things and activities but at times finds difficulty with complicated interactions or rare words, in accordance with reports of recent studies[1][3][4].The fixed-size context requirement of the model and lack of apparent attention mechanisms degrade performance on pictures with numerous important regions or involved interactions[1][2][8], while integration testing preserves stable data exchanges between modules and unit tests guarantee the correctness of pre-processing and model parts. Future work should include attention layers to dynamically focus on relevant image regions during caption generation and investigate transformer-based architectures for better context modelling and scalability[7][8]. Scaling Up Our CNN-LSTM image caption generator attains BLEU-1 scores of up to 0.66 on the Flickr8k test set, indicating its capacity to generate accurate and grammatically fluent captions. The system properly identifies and annotates common items and activities but occasionally falters over complex relationships or rarer terms, as might be expected from results of more recent research. Integration testing ensures an effective flow of data among modules, but unit tests verify that pre-processing and model elements function accurately. Nevertheless, there are still some limitations.

Its use of fixed-size context and the lack of any explicit attention mechanism can be inhibiting for it to perform on images with many salient areas or complex interactions. Future work should include attention layers to dynamically focus on relevant areas of an image during caption generation and explore transformer-based architectures for better context modelling and scalability. It requires augmenting caption quality and overall system usefulness also by including user feedback, multilingual support, and venturing out to larger and more diverse datasets. Larger and more varied datasets, adding multilingual support, and incorporating user feedback are further avenues for improving caption quality and system usefulness.

## REFERENCES

[1] Jambhale, Sangale, et al. "Image caption generator using convolutional neural networks and long short-term memory." *International Research Journal of Modernization in Engineering Technology and Science* 4 (2017): 4664-4670.

[2] https://ijrti.org/papers/IJRTI2208183.pdf

[3] Shinde, O., Gawde, R., & Paradkar, A. (2021). Image caption generation methodologies. *International Research Journal of Engineering and Technology (IRJET)*, *8*(04), 3961-3966.

[4] Liu, Shuang, et al. "Image Captioning Based on Deep Neural Networks." *MATEC Web of Conferences*. Vol. 232. EDP Sciences, 2018.

[5] Padate, Roshni, et al. "Image caption generation using a dual attention mechanism." *Engineering Applications of Artificial Intelligence* 123 (2023): 106112.

[6] Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An overview of image caption generation methods." *Computational intelligence and neuroscience* 2020.1 (2020): 3062706.

[7] Bhuiyan, Ahatesham, et al. "Enhancing image caption generation through context-aware attention mechanism." *Heliyon* 10.17 (2024).

[8] Sailaja, M., et al. "Image caption generator using deep learning." *2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC)*. IEEE, 2022.

[9] Wang, Cheng, et al. "Image captioning with deep bidirectional LSTMs." *Proceedings of the 24th ACM international conference on Multimedia*. 2016.

[10] Amritkar, Chetan, and Vaishali Jabade. "Image caption generation using deep learning technique." *2018 fourth international conference on computing communication control and automation (ICCUBEA)*. IEEE, 2018.

[11] Bhuiyan, Ahatesham, et al. "Enhancing image caption generation through context-aware attention mechanism." *Heliyon* 10.17 (2024).

[12] Image Captioning | Papers With Code https://paperswithcode.com/task/image-captioning

[13] Padate, Roshni, et al. "Image caption generation via improved vision-language pre-training model: perception towards image retrieval." *The Imaging Science Journal* (2025): 1-27.

[14] Tyagi, Shourya, et al. "Novel Advance Image Caption Generation Utilizing Vision Transformer and Generative Adversarial Networks." *Computers* 13.12 (2024): 305.

[15] [15] Verma, Akash, et al. "Automatic image caption generation using deep learning." *Multimedia Tools and Applications* 83.2 (2024): 5309-5325.

[16] Luo, Gaifang, et al. "A thorough review of models, evaluation metrics, and datasets on image captioning." *IET Image Processing* 16.2 (2022): 311-332.

[17] Thobhani, Alaa, et al. "A Survey on Enhancing Image Captioning with Advanced Strategies and Techniques." *Computer Modeling in Engineering & Sciences (CMES)* 142.3 (2025).

[18] [Park, Seokmok, and Joonki Paik. "RefCap: image captioning with referent objects attributes." *Scientific Reports* 13.1 (2023): 21577.

[19] Verma, Akash, et al. "Automatic image caption generation using deep learning." *Multimedia Tools and Applications* 83.2 (2024): 5309-5325.

[20] Kanimozhiselvi, C. S., et al. "Image captioning using deep learning." *2022 International Conference on Computer Communication and Informatics (ICCCI)*. IEEE, 2022.