# A Comprehensive Study on Image Captioning Using CNN and LSTM Frameworks

Himanshu R. Katrojwar
Department of Data Science,
IoT and Cyber Security
G H Raisoni College of Engineering
Nagpur-440016, India
himanshu.katrojwar8104@gmail.com

Sakshi Warkari
Department of Data Science,
IoT and Cyber Security
G H Raisoni College of Engineering
Nagpur-440016, India
sakshi.warkari.ds@ghrce.raisoni.net

Prachi Dhore
Department of Data Science,
IoT and Cyber Security
G H Raisoni College of Engineering
Nagpur-440016, India
prachi.dhore.ds@ghrce.raisoni.net

Sakshi Jiwtode
Department of Data Science,
IoT and Cyber Security
G H Raisoni College of Engineering
Nagpur-440016, India
sakshi.jiwtode.ds@ghrce.raisoni.net

Himanshi Hatwar
Department of Data Science,
IoT and Cyber Security
G H Raisoni College of Engineering
Nagpur-440016, India
himanshi.hatwar.ds@ghrce.raisoni.net

Amit Pandey
IIIT - Hyderabad
Hyderabad, India
amit.pandey@iiith.in

***Abstract*—This study focuses on developing a robust image captioning model that merges convolutional and recurrent neural networks to generate descriptive captions for images. The model leverages the Flickr8k dataset, which provides a diverse range of images with multiple captions, enhancing the learning process. Data preprocessing includes standardizing captions and tokenization to prepare input for training. Feature extraction is achieved using a pre-trained VGG16 model, capturing intricate visual details. The encoder-decoder framework, enhanced by a bidirectional LSTM and an attention mechanism, facilitates contextually relevant caption generation. Evaluation shows effective learning, with BLEU-1 and BLEU-2 scores indicating promising accuracy and coherence. Future work will address model refinements, larger datasets, and advanced attention mechanisms for improved performance. This research highlights the potential of deep learning in generating human-like descriptions and advancing applications in computer vision and NLP.**

*Keywords—Image Captioning, CNN, LSTM, Attention Mechanism, Feature Extraction*

## I. Introduction

Image captioning has become a prominent field of study at the nexus of natural language processing and computer vision in recent years. The main objective of image captioning is to automatically produce textual descriptions for images. This requires knowledge of the relationships and context of the objects in an image in addition to their detection. This work is essential for several applications, such as helping people who are blind or visually impaired, optimizing picture search engines, and boosting multimedia platform content creation.

The field has come a long way thanks to deep learning models like recurrent neural networks (RNNs), which include Long Short-Term Memory (LSTM) networks for sequence generation, and convolutional neural networks (CNNs) for visual feature extraction. One notable achievement was the invention of the Show, Attend, and Tell paradigm, which makes use of attention mechanisms to focus on certain portions of a picture while writing associated captions. When it came to providing descriptions that were more precise and appropriate for the situation, this model shown a noticeable improvement.

Statistics show that image captioning is progressing, as seen by results on benchmark datasets like MSCOCO (Microsoft Common Objects in Context). A measure of the quality of text that has been machine translated from one language to another, the BLEU-1 (Bilingual Evaluation Understudy) score, has improved from roughly 60% in early models to over 75% in more recent approaches. Furthermore, the capacity to produce captions that closely match human-generated descriptions has been substantially improved with the introduction of transformers and attention-based models.

Despite these advancements, image captioning still faces several challenges. These include the generation of captions for images with complex scenes, understanding abstract concepts, and dealing with bias in training datasets. The field continues to evolve with ongoing research focused on improving the accuracy, diversity, and contextual relevance of the generated captions. As the demand for more sophisticated multimedia systems grows, the importance of image captioning as a key technology is expected to increase, driving further innovation and application in various domains.

## II. Literature Survey

Zhenghang Yuan et.al [1] proposed Multi-level attention module for spatial and scale features extraction and attribute encoder with GCN for learning effective attribute features.

The framework enhances remote sensing image captioning using attention and graph convolution. It has achieved superior performance on UCM-captions, Sydney-captions, and RSICD dataset.

Jingqiang Chen et.al [2] proposed Multi-modal attentional mechanism proposed for news image captioning and RNN-based decoder for text and image encoding to generate words. Soheyla Amirian et.al [3] focused on encoder for object and feature extraction in CNN, decoder in neural network for generating natural language sentences and G-MLE and G-GAN strategies for image captioning.

Edy Mulyanto et.al [4] proposed CNN and LSTM models for image caption generation on different languages like English, Chinese, Arabic, and Japanese showing above-average results. It achieved BLEU-1 score of 50.0 and BLEU-3 of 23.9. Shiru Qu et.al [5] proposed LSTM with attention for image captioning achieves state-of-the-art performance. Attention mechanism enhances interpretability and alignment with human intuition. Model combines CNN features with LSTM for semantic sentence generation.

Feng Chen, Songxian Xie et.al [6] addressed Image Captioning challenges in neural network-based models. It uses Template-based methods with attributes, scenes, and objects. Minsi Wang et.al [7] proposed parallel-fusion RNN-LSTM architecture for image caption generation. It uses BLEU, Meteor, and perplexity metrics for sentence generation evaluation.

Xinru Wei et.al [8] focused on image retrieval by dense caption reasoning. It addresses complex image retrieval through structured language descriptions and compares with baseline methods on a large-scale CBIR dataset. Abu Musa Sakib et.al [9] proposed CNN and LSTM layers for image caption generation and feature extraction from images using pre-trained CNN. Further model used text processing for captions, tokenization, and sequence conversion.

Genc Hoxha et.al [10] proposed CNN-RNN framework for remote sensing image captioning and Reviews existing methods in remote sensing image captioning. Binqiang Wang et.al [11] focused on image captioning using memory cells and topics. It uses Template-based methods to fill predefined sentence templates with detected objects, Retrieval-based methods treat captioning as an image-sentence retrieval task and RNN-based methods view sentence generation as a continuous sequence process.

Varsha Kesavan et.al [12] proposed transfer learning approach with pre-trained VGG16 model encoder and attention mechanism, Hard and Soft attention for feature vector representation.

## III. Methodology

This research aims to develop a robust image captioning model that combines convolutional and recurrent neural networks to generate descriptive captions for images. The process involves Data Collection, Data Preprocessing, Feature extraction, model design, and evaluation, as outlined below.

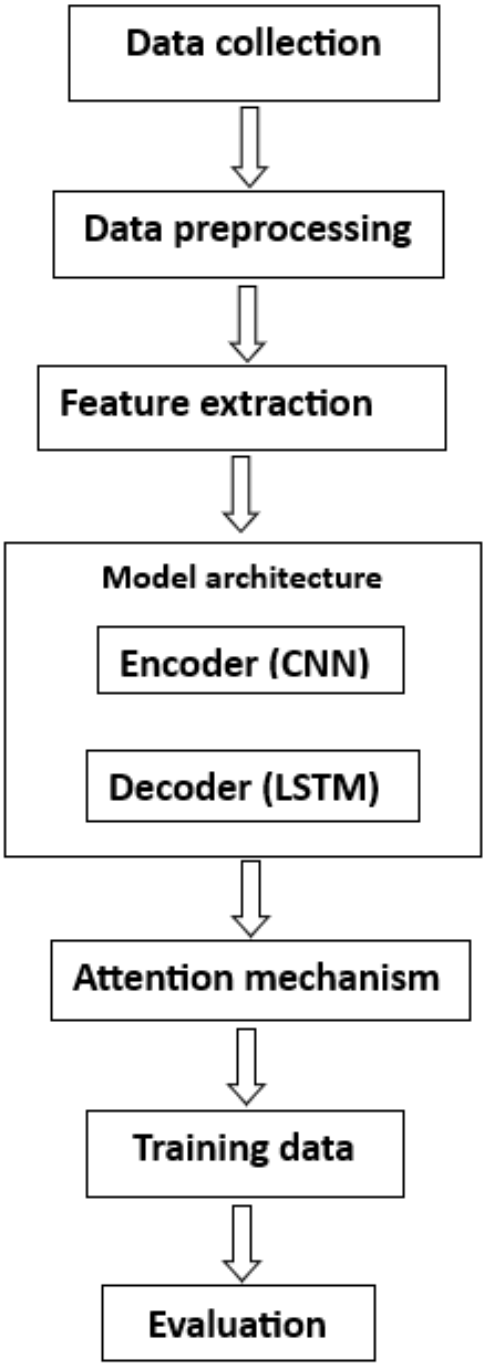


Fig. 1. Overview of Methodology

### A. Data Collection

This research leverages the Flickr8k dataset, which contains 8,091 images, each paired with five distinct captions

offering diverse descriptions of the image content. This multi-caption structure enhances the dataset's ability to capture various perspectives of a scene, making it an ideal resource for training and evaluating image captioning models. The dataset spans a broad spectrum of subjects, including people, animals, and landscapes, enabling the model to generalize across different visual elements and contexts. Furthermore, the high-quality captions are consistently written, facilitating the model's learning of coherent, descriptive, and grammatically accurate caption generation.

### *B. Data Preprocessing*

To prepare the dataset for training the image captioning model, several preprocessing steps were applied to the captions to improve data quality and learning efficiency. Each caption was standardized to lowercase and stripped of special characters, reducing vocabulary size and eliminating variations from capitalization and non-alphabetic characters. This cleaned data was then tokenized by appending unique start (startseq) and end (endseq) tokens to each caption, which clearly marks sequence boundaries for the model, enabling it to predict captions word-by-word and recognize when to terminate.

Additionally, a tokenizer was used to build a vocabulary from the training captions, generating a word-to-index mapping that translates each unique word into a numerical index. This mapping converts textual captions into sequences of numbers, facilitating effective processing and interpretation by the model, ultimately enhancing caption generation accuracy.

### *C. Feature Extraction*

To facilitate effective image content analysis and feature extraction for caption generation, this study employs a pre-trained VGG16 model known for its depth and capacity to capture detailed visual information, including textures, shapes, and colors—qualities essential for complex tasks like image captioning. Utilizing weights pre-trained on the extensive ImageNet dataset, the VGG16 model is adapted for feature extraction by removing its final classification layer, enabling it to output a 4096-dimensional feature vector for each image. Each image is resized to 224x224 pixels to align with VGG16's input specifications, followed by standard normalization to ensure uniform input quality. These extracted feature vectors are stored in a dictionary keyed by image IDs, allowing efficient retrieval during model training and ensuring seamless integration with the caption generation process. This structured feature extraction approach provides a robust foundation for generating accurate, descriptive captions.

### *D. Model Architecture*

The proposed image captioning model is structured as an encoder-decoder framework. The encoder is a modified VGG16 model, and the decoder is a bidirectional Long Short-Term Memory (LSTM) network that generates captions sequentially, word-by-word.

- **Encoder**: The VGG16-based encoder extracts deep visual features from each image, outputting a 4096-dimensional feature vector. To prevent overfitting, a dropout layer with a 0.5 rate follows the feature extraction, and a dense layer further reduces the dimensionality to 256, optimizing compatibility with the decoder.
- **Decoder**: The decoder begins with an embedding layer that maps vocabulary words into dense vector representations, facilitating the LSTM's understanding of syntactic and semantic relationships. The embedded words and encoded image features are processed by a bidirectional LSTM, which captures both forward and backward dependencies for improved coherence in generated captions. Finally, a dense layer with a softmax activation outputs a probability distribution over the vocabulary, enabling word-by-word caption prediction. This architecture ensures that the model generates descriptive and contextually accurate captions.

### *E. Attention Mechanism*

The attention mechanism improves captioning accuracy by allowing the model to focus on relevant image regions. It computes attention scores through dot products between encoder outputs and decoder states, normalizes them via softmax, and uses the resulting weights to create a context vector, influencing the generation of each word.

### *F. Training and Evaluation*

**Training Details:** The model is trained to minimize categorical cross-entropy loss, using a data generator function to create batches of image-caption pairs. Each input consists of image features and partial captions, with the target being the next word. The model uses batch training (batch size = 32) with dropout (rate = 0.5) and the Adam optimizer for better convergence.

**Evaluation Metrics:** Model performance is evaluated using the BLEU score, which compares n-grams between predicted and reference captions. BLEU-1 focuses on unigram matches to assess word-level accuracy, while BLEU-2 evaluates bigram matches to gauge short-phrase coherence. Qualitative analysis manually reviews selected captions for contextual relevance and accuracy.

## IV. Results

The development of the image captioning model yielded significant findings, demonstrating a capability to generate coherent and contextually appropriate captions for images. Leveraging the VGG16 architecture for feature extraction and

an LSTM-based model enhanced with an attention mechanism, the following key outcomes were observed:

- **Model Performance Metrics:**

  Over 50 training epochs, the model exhibited a steady decrease in training loss, indicative of effective learning, culminating in a final loss of 0.4954. The validation loss, reaching 15.6088, reflected the model's generalization potential, albeit with room for improvement. BLEU scores were used to evaluate caption generation proficiency, revealing a BLEU-1 score of 0.3799, which indicated good precision for single-word matches. The BLEU-2 score of 0.1670 suggested a fair ability to create relevant two-word sequences. These metrics demonstrated the model's capability to generate captions with reasonable accuracy and highlighted areas for future enhancements.

- **Qualitative Analysis:**

  The model's predictions generally aligned with human-generated captions, showcasing its capability to understand and describe image content accurately. Examples from test images highlighted its strengths, with most captions reflecting the scene appropriately. However, occasional issues arose with repetitive or overly simplistic descriptions, indicating areas for further refinement. The inclusion of an attention mechanism played a crucial role in improving accuracy by focusing on key image features, thereby enhancing the quality and relevance of the generated captions.

## V. Conclusion

The integration of pre-trained VGG16 for feature extraction and a bidirectional LSTM model with attention for sequence generation successfully demonstrated the feasibility of automated image captioning. The model's ability to comprehend and narrate scenes represents a significant step toward building systems that bridge visual and textual data. While the model achieved respectable BLEU scores, future work could involve refining the model's architecture, using larger and more diverse datasets, and incorporating more sophisticated attention mechanisms to improve both precision and recall in caption generation. Overall, this study underscores the potential of deep learning approaches in generating human-like descriptions from images, contributing to advancements in computer vision and natural language processing applications.

## References


[1] Z. Yuan, X. Li and Q. Wang, "Exploring Multi-Level Attention and Semantic Relationship for Remote Sensing Image Captioning," in IEEE Access, vol. 8, pp. 2608-2620, 2020, doi: 10.1109/ACCESS.2019.2962195

[2] J. Chen and H. Zhuge, "News Image Captioning Based on Text Summarization Using Image as Query," 2019 15th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, 2019, pp. 123-126, doi: 10.1109/SKG49510.2019.00029.

[3] S. Amirian, K. Rasheed, T. R. Taha and H. R. Arabnia, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap," in IEEE Access, vol. 8, pp. 218386-218400, 2020, doi: 10.1109/ACCESS.2020.3042484.

[4] E. Mulyanto, E. I. Setiawan, E. M. Yuniarno and M. H. Purnomo, "Automatic Indonesian Image Caption Generation using CNN-LSTM Model and FEEH-ID Dataset," 2019 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA), Tianjin, China, 2019, pp. 1-5, doi: 10.1109/CIVEMSA45640.2019.9071632.

[5] S. Qu, Y. Xi and S. Ding, "Visual attention based on long-short term memory model for image caption generation," 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, China, 2017, pp. 4789-4794, doi: 10.1109/CCDC.2017.7979342.

[6] F. Chen, S. Xie, X. Li, S. Li, J. Tang and T. Wang, "What Topics Do Images Say: A Neural Image Captioning Model with Topic Representation," 2019 IEEE International Conference on Multimedia Expo Workshops (ICMEW), Shanghai, China, 2019, pp. 447-452, doi: 10.1109/ICMEW.2019.00083.

[7] M. Wang, L. Song, X. Yang and C. Luo, "A parallel-fusion RNN-LSTM architecture for image caption generation," 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2016, pp. 4448-4452, doi: 10.1109/ICIP.2016.7533201.

[8] X. Wei, Y. Qi, J. Liu and F. Liu, "Image retrieval by dense caption reasoning," 2017 IEEE Visual Communications and Image Processing (VCIP), St. Petersburg, FL, USA, 2017, pp. 1-4, doi: 10.1109/VCIP.2017.8305157.

[9] Sakib AM, Mukta SA, Hossain MY. Automated Image Captioning System, April 2024, License CC0, DOI: 10.13140/RG.2.2.24966.79689.

[10] G. Hoxha, F. Melgani and J. Slaghenauffi, "A New CNN-RNN Framework For Remote Sensing Image Captioning," 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), Tunis, Tunisia, 2020, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.

[11] B. Wang, X. Zheng, B. Qu and X. Lu, "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 256-270, 2020, doi: 10.1109/JSTARS.2019.2959208.

[12] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," 2019 Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2019, pp. 1-6, doi: 10.1109/GCAT47503.2019.8978293.