INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

SYSTEMATIC REVIEW ON IMAGE CAPTIONING USING DEEP LEARNING TECHNIQUES

[1] Ms. Jinal V. Purohit, [2]Dr. Priti Patel

[1]Assistant Professor, [2] Assistant Professor
[1]Department of Computer Science
[1] Vanita Vishram Women's University, Surat, India

Abstract: Image captioning is method of generating descriptions for images by integrating computer vision and natural language processing techniques. This paper presents a systemic review of image captioning using deep learning approaches, like CNN-RNN, CNN-CNN, CNN-LSTM reinforcement-based, and transformer-based models. Datasets such as MS COCO, Flickr8k, Flickr30k, VizWiz and Visual Genome, along with evaluation metrics like BLEU, ROUGE, and METEOR, provide the foundation for training and benchmarking. Significant work is done in the field but still some challenges include handling complex scenes, ensuring contextual accuracy, reducing delusions, and improving caption diversity needs to be addressed. Future directions highlight the importance of vision-language pre-training, improved evaluation tools, and scalable models, with applications in accessibility for the visually impaired, social media content monitoring, and digital learning platforms.

Index Terms **- Image Captioning, Deep Learning, CNN, RNN, LSTM, Encoder-Decoder.**

I. INTRODUCTION

Image captioning is a process that links computer vision and natural language processing by generating descriptive sentences for images. It empowers machines to move beyond object recognition toward semantic understanding of scenes, thereby imitating human cognitive abilities. With applications ranging from diagnosing and treating a range of health condition in medical, to improve the customer shopping experience in retail and marketing field, assistive technologies for the visually impaired and automated content generation to enhanced search engines, and intelligent Internet of Things (IoT) systems, image captioning has emerged as a crucial research area.
Traditional systems, such as template-based and retrieval-based models can provide pertinent captions for images, they often lack diversity and creativity, since they rely on established templates, making it difficult to generate new or contextually appropriate descriptions for unseen query images [1]. Nowadays most Image captioning systems adopt an encoder–decoder framework, where an encoder such as a convolutional neural network (CNN) or a visual transformer extracts visual features, and a decoder such as a recurrent neural network (RNN), long short-term memory (LSTM), or transformer generates descriptive text [3]. Training is typically supervised on large-scale paired datasets like MS COCO, Flickr8k, and Flickr30k, with optimization guided by evaluation metrics such as the Consensus-based Image Description Evaluation (CIDEr) score. [3]

Broadly, captioning methods can be divided into three categories: retrieval-based methods, which select captions from an existing database; template-based methods, which rely on predefined sentence structures; and deep learning-based methods, which directly generate captions by mapping image features to natural language. Recent advances in deep learning have significantly improved captioning performance. CNNs and RNNs contribute to robust feature extraction and sequence modelling, while attention mechanisms allow models to focus on salient image regions. Moreover, transformers and generative adversarial networks (GANs) have further enhanced contextual coherence, diversity, and naturalness of captions. [4]

Nevertheless, several challenges remain unresolved. Capturing the complexity of visual scenes with multiple objects and interactions remains difficult. Ensuring linguistic fluency, avoiding generic or repetitive captions, and achieving deeper contextual reasoning are ongoing issues. To address these gaps, a review in [4] provides a comprehensive survey of image captioning. Authors contributions are fivefold: (i) development of taxonomies for natural and medical image captioning from the perspective of training paradigms and deep models, (ii) simulation of captioning as a process of seeing, focusing, and telling, (iii) review of commonly used datasets, metrics, and loss functions, (iv) qualitative and quantitative comparisons of representative captioning models, and (v) identification of open challenges and future research directions[4].

II. IMAGE CAPTIONING EVOLUTION

Image captioning algorithms are typically divided into three categories. The first category, is the retrieval-based methods, which first retrieves the closest matching images, and then transfer their descriptions as the captions of the query images [5]. These methods can produce grammatically correct sentences but cannot adjust the captions according to the new image. The second category is template-based methods to generate descriptions with predefined syntactic rules and slit sentences into several parts [6]. These methods first take advantage of several classifiers to recognize the objects, as well as their attributes and relationships in an image, and then use a rigid sentence template to form a complete sentence. Though it can generate a new sentence, these methods either cannot express the visual context

Figure 1 progressive development of image captioning system. [1]

correctly or generate flexible and meaningful sentences. Most recent works fall into the third category called neural network-based methods Inspired by machine learning's encoder-decoder architecture [7], recent years most image captioning methods employ a Convolutional Neural Network (CNN) as the encoder and a Recurrent Neural. Network (RNN) as the decoder, especially Long Short-Term Memory (LSTM) [8] to generate captions [9], with the objective to maximize the likelihood of a sentence given the visual features of an image. Some methods are using CNN as the decoder and the reinforcement learning as the decision-making network.
In this paper, we explore the image captioning methods of deep learning methods like encoder-decoder method, CNN and RNN, LSTM.

III. DEEP LEARNING -IMAGE CAPTIONING METHODS

3.1 Encoder-Decoder Architecture-Based Image captioning

The Encoder-Decoder framework for image captioning uses an encoder (typically a Convolutional Neural Network (CNN)) to process an input image and extract its visual features into a compact representation, and a decoder (often a Recurrent Neural Network (RNN) like LSTM) to generate a human-readable text caption by iteratively producing words based on the encoded image features. This architecture is a powerful sequence-to-sequence model used for tasks that map one type of data to another, like images to text.

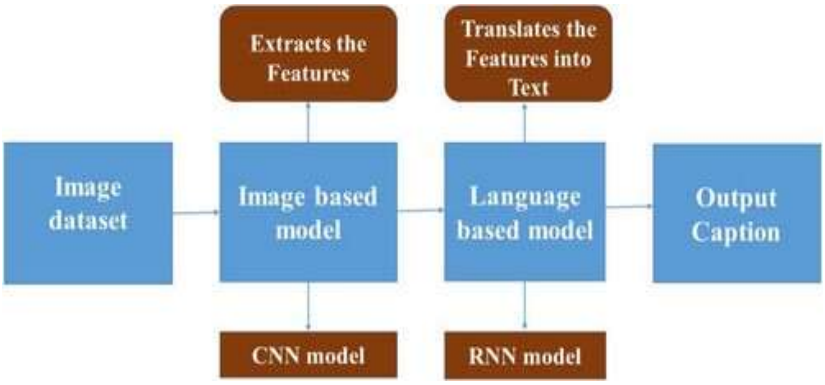

Figure 2 Encoder-Decoder model [10]

The model as shown in figure can categorized in two phases vice;
Encoder (Visual Feature Extraction):

A pre-trained CNN (like VGG, ResNet, or InceptionNet) takes the input image as its input. The CNN processes the image through layers of convolution and pooling, progressively extracting high-level visual features. The final output of the encoder is a feature vector or feature map that represents the semantic content of the image
Decoder (Text Generation):

The encoded image features are then fed into the decoder, which is usually an RNN-based model like a Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU). The decoder generates the caption one word at a time. At each step, the decoder uses the image features and the previously generated word to predict the next word in the sequence. This process continues until a special end-of-sentence token is generated, signalling the completion of the caption.

3.2 Attention mechanisms

Attention mechanisms in deep learning-based image captioning improve accuracy by allowing the model to dynamically focus on relevant parts of an image as it generates each word of the caption, rather than treating the entire image uniformly.

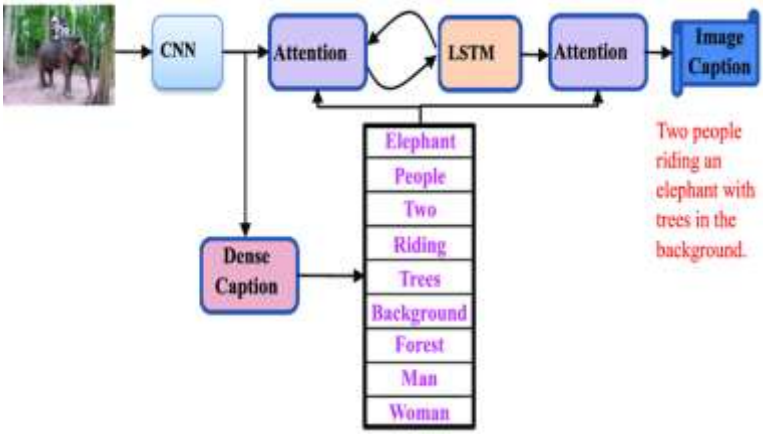

Figure 3 Attention mechanisms [11]

The model divides the image in to regions, and for each word being generated, the attention mechanism calculates attention weights to highlight the most significant image regions. This ensures that, for example, when the model generates the word "dog," it pays more attention

to the region of the image containing the dog, leading to more relevant and descriptive captions. It processes in following steps like Image Encoding, Dividing the Image, Query Generation, Calculating Attention Weights, Weighted Combination, Contextual Captioning.

3.3 Transformers in Image Captioning

Transformers have transformed image captioning by replacing conventional CNN–RNN architectures with attention-based models, enabling more accurate and contextually relevant descriptions [15]. In typical implementations, a Vision Transformer (ViT) encoder converts the image into vector embedding, which are processed through self-attention to capture global contextual relationships. [13] The decoder then

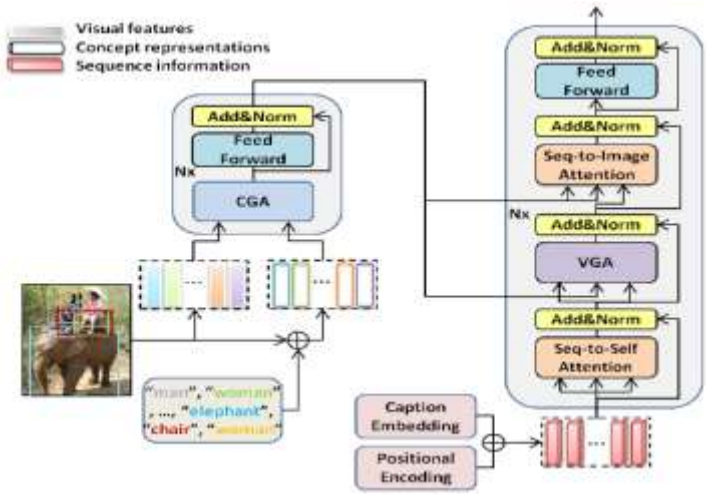

Figure 4 Transformers in Image Captioning [12]

employs cross-attention to align textual generation with the most relevant image features, producing captions word by word. [14] This mechanism allows the model to dynamically attend to specific image regions (e.g., focusing on an object such as a *dog* when generating the corresponding token), thereby enhancing caption precision and coherence.

IV. DATA-SETS

There are various benchmark datasets available for Computer Vision and natural language processing tasks, specifically for image captioning and object recognition. The MS COCO dataset contains over 3,30,000 images with five human-annotated captions per image, the ImageNet dataset is used for object classification and recognition, containing over 1,281,167 images labelled across 20,000 categories. The Flickr datasets are extensively used for image captioning tasks. The Flickr30k dataset contains 30,000 images with five descriptive captions for each image. The Pascal VOC (Visual Object Classes) dataset is designed for object detection, segmentation, and classification, containing over 11,530 images with 20 object categories. [1]

V. EVALUATION METRICES

Evaluating the trained model is quiet difficult task in image captioning for this purpose various evaluation matrices are created.
In [16] the authors study mostly uses the degree of matching between the caption sentence and the reference sentence to evaluate the pros and cons of the generation results. The commonly used methods include BLEU, METEOR, ROUGE, CIDEr and SPICE these five measurement indicators. Among them, BLEU and METEOR are derived from machine translation, ROUGE is derived from text abstraction, and CIDEr and SPICE are specific indicators based on image captioning. [16]
BLEU is widely used in the evaluation of image annotation results, which is based on the n-gram precision. The principle of the BLEU measure is to calculate the distance between the evaluated and the reference sentences. BLEU method tends to give the higher score when the caption is closest to the length of the reference statement.

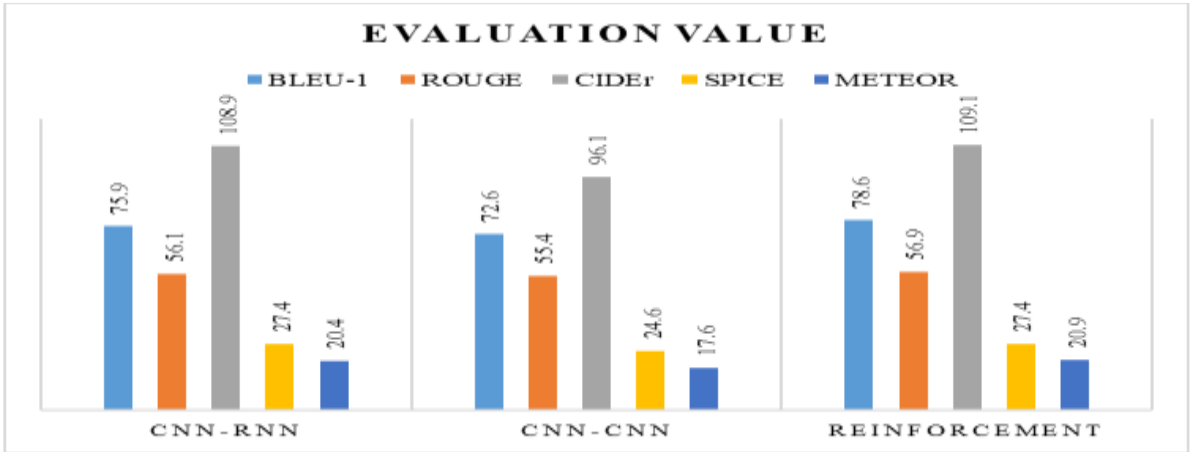

Figure 5 Evaluation Index of three methods [16]

ROUGE is an automatic evaluation standard designed to evaluate text summarization algorithms. There are three evaluation criteria, ROUGE-N, ROUGE-L, and ROUGE-S. ROUGE-N is based on the given sentence to be evaluated, which calculates a simple n-tuple recall for all reference statements: ROUGE-L is based on the largest common sequence (LCS) calculating the recall. ROUGE-S calculates recall based on co-occurrence statistics of skip-bigram between reference text description and prediction text description.
CIDEr is the special method which is provided for the image captioning work. It measures consensus in image captioning by performing a term frequency inverse document frequency (tf-idf) for each n-gram. Studies have shown that the match between CIDEr and human consensus is better than other evaluation criteria.
METEOR is based on the harmonic mean of unigram precision and recall, but the weight of the recall is higher than the accuracy. It is highly relevant to human judgment and differs from the BLEU in that it is not only in the entire set, but also in the sentence and segmentation levels, and it has a high correlation with human judgment.
SPICE evaluates the quality of image captions by converting the generated description sentences and reference sentences into graph-based semantic representations, namely "scene graphs". The scene graphs extract lexical and syntactic information in natural language and explicitly represents the objects, attributes, and relationships contained in the image

In Fig.5, [16] have shown the best results of the above three methods for five evaluation metrics. The authors confirm that both the CNN-RNN based and the Reinforcement based methods can get the better performance than the CNN-CNN based framework, which greatly improves the training speed without seriously affecting the accuracy.

Table 1: Evaluation Metrics for Image Captioning Model

Name	Value
BLEU Score	0.27
Semantic Similarity Score	0.89
ROUGE-1 Score	0.38
ROUGE-2 Score	0.07
ROUGE-L Score	0.37

In [2] the authors have presented a table that shows the results of various metrices for image captioning. The BLEU score (0.2709) shows moderate overlap with reference captions, reflecting fair linguistic quality. The Semantic Similarity score (0.8898) indicates strong semantic relevance, suggesting the model effectively captures the meaning of images. ROUGE scores highlight n-gram and sequence overlaps: ROUGE-1 (0.3821) shows reasonable unigram matches, ROUGE-2 (0.0733) reveals weak bigram agreement, and ROUGE-L (0.3708) indicates moderate longest common subsequence overlap. Overall, the model produces semantically accurate captions but with limited lexical and structural alignment to references.

VI. COMPARATIVE ANALYSIS OF IMAGE CAPTIONING TECHNIQUES

The comparison of various image captioning techniques is presented in Table 2. The table highlights different models, datasets, methodologies, and their performance, providing a clear overview of existing approaches.

Sr. No.	Author(s)	Year	Model Type	Dataset	Key Techniques	Performance Insight	Limitations
1	Huda & Basir	2025	Transformer + MLLM	MS COCO	Multimodal LLMs	High semantic accuracy	High computational cost
2	Shan-E-Fatima	2024	CNN + LSTM	Flickr8k	Basic deep learning	Moderate BLEU	Limited generalization
3	Bhosale et al.	2025	Review	Multiple	Comparative study	Identifies trends	No experimental results
4	Xu et al.	2023	Attention-based	MS COCO	Visual attention	Improved context	Still misses fine details
5	Devlin et al.	2015	Language Model	-	Caption refinement	Improved fluency	Weak visual grounding
6	Fang et al.	2015	CNN + RNN	MS COCO	Visual concept detection	Better object recognition	Limited sentence quality
7	Cho et al.	2014	Encoder-Decoder	-	RNN seq2seq	Foundation model	Not image-specific

8	Hochreiter & Schmidhuber	1997	LSTM	-	Memory units	Handles sequences well	No visual understanding
9	Karpathy & Li	2015	CNN + RNN	Flickr30k	Alignment model	Good image-text mapping	Limited scalability
10	Kavitha & Karpagam	2025	Hybrid (CNN+LSTM+GRU)	MS COCO	Beam search optimization	High BLEU, CIDEr	Complex architecture
11	Liu & Xu	2020	Attention-based	MS COCO	Adaptive attention	Better semantics	Moderate complexity
12	Li et al.	-	Transformer	-	Boosted transformer	Improved captioning	Limited validation
13	Dosovitskiy et al.	2021	Vision Transformer	ImageNet	Patch-based learning	Strong feature extraction	Needs large data
14	Herdade et al.	2019	Object-based	MS COCO	Region-based captioning	Better object relations	Computational cost
15	Vaswani et al.	2017	Transformer	-	Self-attention	State-of-the-art NLP	Not image-specific initially
16	Liu et al.	-	DNN-based	-	Deep neural networks	General captioning	Lack of

VII. FUTURE DIRECTIONS

The ability to guide the captioning process to emphasize specific visual aspects or generate a desired caption style. Making the models decision-making process more transparent so users can understand why a particular caption was generated. Moving beyond simple descriptions to capture the broader context, subtle relationships, and nuances within an image. Improving the ability to correctly combine multiple visual elements and their interactions to form a comprehensive caption. Addressing the issue where models generate descriptions for objects that are not actually present in the image. Ensuring that all relevant objects and actions are included in the generated descriptions. Generating a wider range of unique and natural captions for the same image, rather than relying on repetitive phrasing.

VIII. CONCLUSION

Deep learning models effectively caption images by combining Convolutional Neural Networks (CNNs) for feature extraction and Recurrent Neural Networks (RNNs) or Transformers for language generation, enabling machines to interpret and describe the visual world. These systems have significantly advanced availability, content generation, and assistive technologies by linking the visual and textual data. While deep learning image captioning has shown great success, future work focuses some of the challenges like struggle to describe complex, cluttered scenes or objects not present in their training data, limiting their understanding of novel visual content, improving its accuracy, contextual awareness, and ability to handle complex scenes and diverse inputs, with a strong emphasis on developing more advanced architectures like transformers and incorporating attention mechanisms.

IX. REFERENCES

[1] Huda Diab Abdulgalil and Basir, O.A. (2025). *Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to multimodal large language models*. Natural Language Processing Journal, 12, 100159. https://doi.org/10.1016/j.nlp.2025.100159
[2] Shan-E-Fatima (2024). *Image caption generation using deep learning algorithm*. Educational Administration: Theory and Practice, 30(5), pp. 8118–8128. https://doi.org/10.53555/kuey.v30i5.4311
[3] Bhosale, C.S., Salve, P. and Shirsath, V. (2025). *A review of image captioning techniques: Types, deep learning advancements, and limitations*. Cureus Journal of Computer Science, 2, es44389-024-01573-w. https://doi.org/10.7759/s44389-024-01573-w
[4] Xu, L., Tang, Q., Lv, J., Zheng, B., Zeng, X. and Li, W. (2023). *Deep image captioning: A review of methods, trends and future challenges*. Neurocomputing, 546, 126287. https://doi.org/10.1016/j.neucom.2023.126287
[5] Devlin, J. *et al.* (2015). *Language models for image captioning: The quirks and what works*. Computer Science.
[6] Fang, H. *et al.* (2015). *From captions to visual concepts and back*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1473–1482.
[7] Cho, K. *et al.* (2014). *Learning phrase representations using RNN encoder-decoder for statistical machine translation*. Computer Science.
[8] Hochreiter, S. and Schmidhuber, J. (1997). *Long short-term memory*. Neural Computation, 9(8), pp. 1735–1780.
[9] Karpathy, A. and Li, F.F. (2015). *Deep visual-semantic alignments for generating image descriptions*. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3128–3137.
[10] Kavitha, P.V. and Karpagam, V. (2025). *Image captioning deep learning model using ResNet50 encoder and hybrid LSTM–GRU decoder optimized with beam search*. Automatika, 66(3), pp. 394–410. https://doi.org/10.1080/00051144.2025.2485695

[11] Liu, X. and Xu, Q. (2020). *Adaptive attention-based high-level semantic introduction for image caption*. ACM Transactions on Multimedia Computing, Communications, and Applications, 16, pp. 1–22. https://doi.org/10.1145/3409388
[12] Li, J., Yao, P., Guo, L. and Zhang, W. (n.d.). *Boosted transformer for image captioning*.
[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T. and Houlsby, N. (2021). *An image is worth 16x16 words: Transformers for image recognition at scale*. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/2010.11929
[14] Herdade, S., Kappeler, A., Boakye, K. and Soares, J. (2019). *Image captioning: Transforming objects into words*. Advances in Neural Information Processing Systems (NeurIPS), 32.
[15] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N. and Polosukhin, I. (2017). *Attention is all you need*. Advances in Neural Information Processing Systems (NeurIPS), 30.
[16] Liu, S., Bai, L., Hu, Y. and Wang, H. (n.d.). *Image captioning based on deep neural networks*. College of Systems Engineering, National University of Defense Technology, Changsha, China.