International Journal of
Advanced Science Computing and Engineering

Journal Homepage www.ijasce.org/index.php/ijasce

Image Caption Generator Using Bahdanau Attention Mechanism

Nikhita B Gowda [a], Vaishnavi [a,*], Avin Skanda B N [a], Rohan M [a], Pratheek V Raikar [a]

[a] *Department of Computer Science & Engineering JSS Science and Technology University Mysuru, India*

Corresponding author: [*]vaishud_1393@gmail.com

***Abstract*—This project proposes a sophisticated image captioning system developed using an encoder-decoder framework bolstered with an attention mechanism. The system generates contextually appropriate text descriptions by dynamically weighting relevant image regions with CNNs for feature extraction and RNNs with attention layers. The model shows significant improvement on the Flickr8k dataset, as measured by BLEU. The study examines the use of such systems across domains, including assistive devices and automated indexing, and proposes employing transformer-based attention methods in future upgrades. The development of an image captioning system with an attention mechanism is a key advancement in computer vision and natural language processing. This mechanism helps the model focus on relevant image parts when generating words, improving contextual relevance and semantic accuracy. It aligns visual features with language more effectively, producing captions similar to human descriptions. The model employs teacher forcing during training to accelerate learning and improve fluency. Standard metrics like BLEU evaluate performance and compare models. Inspired by works like "Show, Attend and Tell," attention bridges image features and language. Attention-based captioning can aid visually impaired users, enable content indexing, and improve human–computer interaction. Future research will likely scale models to larger datasets and enhance generalization across diverse scenes.**

***Keywords*—Image captioning; Bahdanau attention; CNN-LSTM; BLEU score; Flickr8k.**

Manuscript received 15 Apr. 2025; revised 12 Jun. 2021; accepted 24 Oct. 2025. Date of publication 30 Dec. 2025.

I. INTRODUCTION

Image captioning systems integrate computer vision and natural language processing to automatically generate detailed, meaningful textual descriptions of visual content [1]. The progress of these systems has been significantly accelerated by advances in deep learning and the proliferation of large-scale datasets, enabling the production of multilingual captions beyond English. Although these sophisticated models simulate human-like understanding and description of images, they require complex algorithms and substantial computational resources to operate efficiently and precisely. Recent developments in this domain have ushered in the era of automatic captioning, thereby creating numerous promising opportunities across diverse sectors, including social media and accessibility services [2].

A diverse array of techniques has emerged for generating image captions, each offering distinct advantages. Among the primary approaches are the integration of Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs), which provides a powerful framework for understanding both the spatial and temporal dimensions of imagery [3]. Attention-based models, such as the innovative Show, Attend, and Tell (SAT) architecture, allow for a more nuanced interpretation of images by focusing on specific regions while generating descriptive text [4].

In addition, DenseCap stands out by generating object-centric captions that highlight individual elements within a scene [5], thereby enhancing the descriptive quality of the produced captions. Further enhancing these techniques, reinforcement learning has been applied to optimize evaluation metrics, aiming for even higher standards of accuracy and relevance in the generated captions.

Nevertheless, the field grapples with ambiguities, largely arising from a heavy reliance on training data and on the evaluation metrics used. Additionally, the computational complexity of these methods can pose significant challenges, limiting their accessibility and practical implementation in real-world applications.

The proposed image captioning system is a meticulously structured program that proceeds through data collection and culminates in the use of the well-known MSCOCO and Flickr8k datasets [6], [7]. This innovative project harnesses the power of a pre-trained Convolutional Neural Network (CNN), such as VGG, serving as an encoder to adeptly extract intricate visual features from images [8], [9] . These features

are then seamlessly transferred to an LSTM-based decoder, which employs a sophisticated attention mechanism. This mechanism enables the model to dynamically focus on different regions of the image, ensuring nuanced, contextually relevant caption generation.

Throughout the training process, the model is optimized using cross-entropy loss and employs teachers forcing to improve learning efficiency. During inference, strategies such as beam search are applied to generate the most accurate predictions. To thoroughly assess the effectiveness and coherence of the captions produced, the authors use BLEU scores, a quantitative metric, alongside qualitative evaluations by human judges, thereby ensuring a comprehensive analysis of output quality.

II. Materials and Methods

A. Existing System

Traditional image captioning systems rely on CNN-RNN architectures[10], with later improvements incorporating attention mechanisms [11], [12]. Recent work by Zhang et al. [13] demonstrated the efficacy of AoA (Attention on Attention) modules, while Vinyals et al. [14] established encoder-decoder frameworks as a baseline. However, these systems face challenges in computational efficiency and contextual accuracy.

B. Proposed System

The proposed image captioning system utilizes a state-of-the-art convolutional neural network, specifically VGG16, as its encoder to extract rich visual features from images. This encoder is paired with a sophisticated long short-term memory (LSTM) decoder that incorporates Bahdanau attention mechanisms [15]. This design enables the model to dynamically assign varying degrees of importance to image regions during captioning, thereby enhancing contextual relevance and the accuracy of the generated descriptions. For training, the system leverages the comprehensive Flickr8k dataset, employs teacher forcing, and evaluates performance using BLEU, a widely recognized metric for measuring the quality of generated text against reference captions.

C. System Design

Figure 1 depicts the proposed image captioning architecture, in which a convolutional neural network (CNN) extracts image features, and a long short-term memory (LSTM) generates a sequence from these features. The structure is encoder-decoder with an attention mechanism, and evaluation is performed using BLEU.

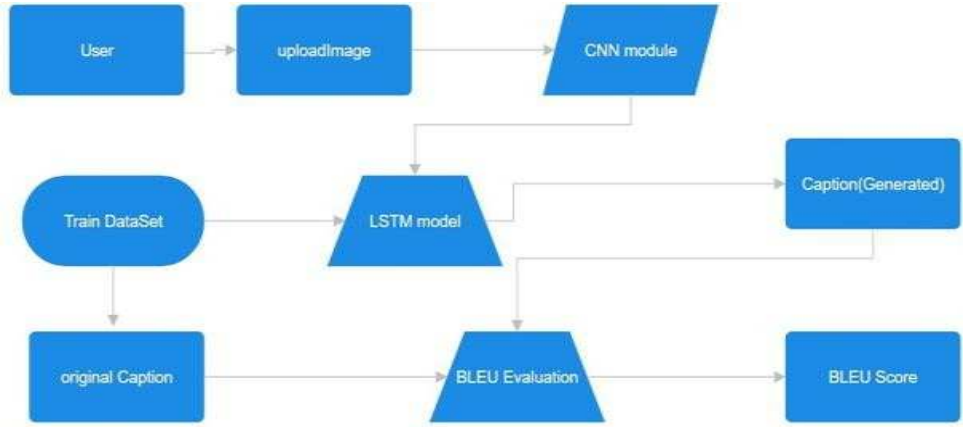

Fig. 1 Image Caption Generator Architecture

1) User Interaction: The first step consists of a user submitting a picture. It is easy to submit images for captions using the interface for a major in elementary education. Based on the pre-service teachers' survey results, exploratory factor analysis, reliability analysis, and confirmatory factor analysis were conducted to validate the items.

2) CNN Model for Feature Extraction: This project uses a pre-trained Convolutional Neural Network (CNN), such as the VGG16 model, to extract deep visual features from the uploaded image. The image features, organized at various levels of the CNN, provide abstract concepts. They form a dense representation of the image. Further, they capture spatial information necessary for understanding the image context. A CNN produces a fixed-length vector that summarizes the image.

3) LSTM Model for Caption Generation: The visual features extracted by the CNN module are provided as input into the LSTM network, which functions as a decoder. The LSTM generates a word sequence (the caption) by modeling the dependencies between visual features and words. The model is trained on the Flickr8k dataset, which consists of image-caption pairs, to learn associations between visual features and textual descriptions and to map visual information to language.

4) Attention Mechanism: To improve caption accuracy, an attention mechanism is employed. The model dynamically adjusts its focus across image regions during the generation of each word in the caption, thereby improving its ability to produce contextually appropriate descriptions. The attention mechanism assigns varying weights to image regions, ensuring that the model places greater emphasis on specific parts during word prediction.

5) Training Process: The system is trained by utilizing a dataset comprising images paired with corresponding human-

annotated captions. The LSTM network is trained to predict the subsequent word in the sequence by leveraging image features and the context provided by previously generated words.

6) Evaluation using BLEU metrics: Once captions are generated, the system evaluates itself using the BLEU metric. BLEU measures the similarity between the generated caption and the human-generated reference and assigns a score to the model's prediction. A high BLEU score indicates that the model-generated caption is similar to the ground-truth caption.

7) Generated Caption and BLEU Score: Once the caption is generated, it is displayed to the user. The BLEU score is used to assess the accuracy of a caption and to provide feedback for improving the model's predictions.

D. Implementation Steps

The project deals with the encoder-decoder attention mechanism-based image captioning project. The project employs several techniques, including an RNN (LSTM), to generate sentences. CNNs are also used to extract various features from images. An attention mechanism is a mechanism that humans employ. They tend to focus on specific regions of the image while ignoring the remainder. Thus, an attention mechanism reduces image noise and improves accuracy. Figure 2 depicts the project implementation, indicating the sequence in which the tasks are to be performed.

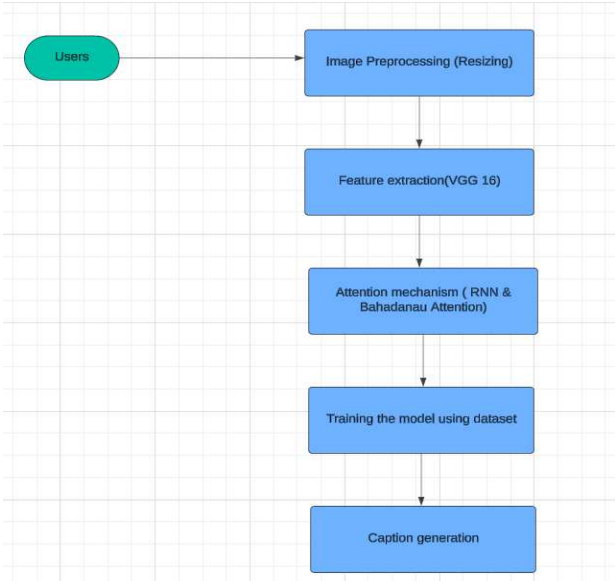

Fig. 2 Workflow of the Project

1) Step 1: Import Libraries: TensorFlow is used for model creation and training, with optional GPU acceleration available through Google Colab or Kaggle.

2) Step 2: Data Loading and Preprocessing: The dataset is loaded, and captions are cleaned by removing punctuation, single characters, and numbers. Each caption is tagged with <start> and <end> markers, and the dataset is batched for training. <start> marker: Signals the model to begin generating or processing a sequence, helping it learn the context from the very first word. <end> marker: Informs the model when the caption or sequence is complete, so it can stop predicting further words.

3) Step 3: Model Setup and Feature Extraction: The VGG16 model is employed to extract image features, with all images resized to 224x224 pixels. These extracted features are saved in .npy format and used as input for the encoder. The model architecture includes an LSTM-based decoder integrated with Bahdanau attention. An 80-20 split is applied to create training and validation datasets. With attention, the model can attend to different parts of the input while making predictions. It doesn't assign equal importance to every token. This technique enhances the model's ability to capture long-distance dependencies and recognize contextual connections among input elements. The Bahdanau attention mechanism is designed to enable the model to look at specific regions in the image when generating each word in the caption. Based on the predicted word, it can quickly shift focus to retrieve information from the image as needed. Consequently, the descriptions are more accurate and contextual. By checking the decoder's hidden state with the encoded image features, we calculate the attention weights. Encoder: An encoder is a pre-trained VGG16 model that extracts key features from an image. Decoder: An LSTM-based recurrent neural network with Bahdanau attention that focuses on key regions of the image during caption generation [15]. The parameters required for the encoder-decoder model are defined, and the loss function and optimizers are configured for training.

4) Step 4: Model Training: The model is trained using teacher forcing, where the target word is passed as the next input to the decoder. The loss is monitored and plotted over epochs for optimization.

5) Step 5: BLEU Evaluation BLEU (Bilingual Evaluation Understudy) is a metric for assessing the quality of machine-generated translations by comparing them with one or more reference translations. It is widely used in natural language processing and machine translation tasks.

III. Results and Discussion

A. Literature Review and Analysis

Recent advancements in image captioning have been driven by innovations in encoder-decoder architectures, attention mechanisms, and domain-specific applications. Key contributions include:

1) Foundational Architecture: Vinyals et al. [14] pioneered the CNN-RNN encoder-decoder framework, achieving state-of-the-art results on datasets like Flickr30k. Their work demonstrated the viability of deep learning for caption generation, particularly in aiding visually impaired users.

2) Attention Mechanisms: Zhang et al. [13] introduced CATANIC, a framework combining InceptionV3, Transformer, and an Attention-on-Attention (AoA) module. This refined focus on image regions and word relevance improves BLEU scores on Flickr8k. Similarly, Bahdanau's local attention [15] reduced computational costs by dynamically weighting image regions during caption generation.

3) Hybrid and Specialized Models: A parallel-fusion technique for RNN-LSTM hybrids, enhancing short- and long-term dependency modeling [16]. Sheng and Moens [5] [17]tailored encoder-decoder models for ancient artworks, showing that domain-specific captioning improves accuracy. ResNet is integrated with LSTM for assistive applications, addressing challenges like overfitting in RNNs [12].

4) Decoding and Evaluation: Chowdhury and Caragea [18] optimized beam search for LSTM/RNN decoders, increasing caption diversity and fluency. BLEU metrics, as used by Kudugunta et al. [18] remain critical for benchmarking against human references.

5) Early Innovations: Pan et al. [19] pioneered keyword-image attribute correlations in 2004, achieving 45% accuracy improvements on Corel datasets. This is later extended this by incorporating deep learning frameworks for cross-language and video annotation [20].

6) Object-Centric Approaches: Kumar et al. [21] enhanced accuracy using Regional Object Detectors (RODe) for feature extraction, emphasizing object-level context in Flickr8k.

These studies collectively underscore the importance of attention mechanisms, hybrid architectures, and domain adaptation in advancing image captioning. This project builds on these foundations by integrating Bahdanau attention into a VGG16-LSTM framework to improve contextual accuracy.

B. Results and Snapshots

Given an image path from the test data, our model predicts a caption.

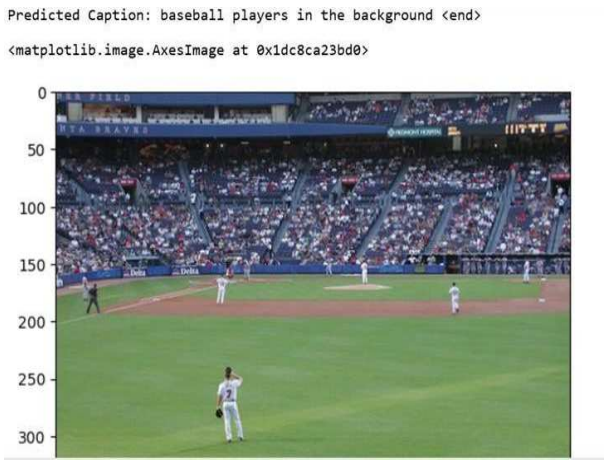

Fig. 3 A team playing baseball.

In Figure 3, the input image featured a team of baseball players on a field, and the model generated the caption "A team playing baseball". Figure 4 displayed a group of trekkers seated on the snowy edge of a mountain, for which the caption was "A group of people sitting in snow". In Figure 5, the image showed a man climbing a snow-covered mountain, and the model produced the caption "A man climbing a snow mountain".

Fig. 4 A group of people sitting in snow

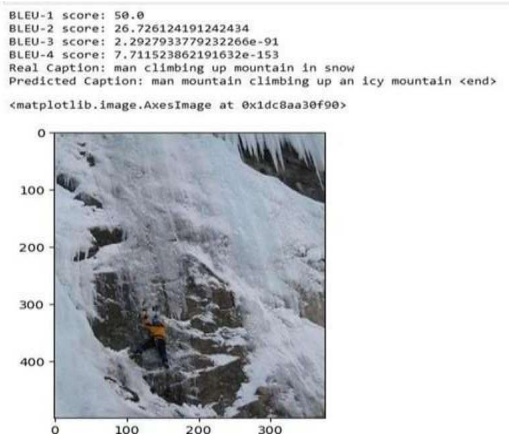

Fig. 5 A man climbing snow mountains.

Fig. 6 A child playing in water

In Figure 6, the input image depicts a young girl playing in water, and the model generated the caption "A child playing in water". These results demonstrate the model's ability to interpret diverse visual scenes and generate relevant, context-aware captions. The consistency and semantic correctness across various scenarios highlight the effectiveness of the proposed image captioning approach.

IV. Conclusion

In sum, the development of an image captioning system incorporating an attention mechanism represents a significant advancement in computer vision and natural language processing. The attention mechanism enables the model to dynamically focus on relevant regions of the image as it generates each word in the caption, thereby enhancing contextual relevance and semantic precision. This aligns visual features with linguistic tokens more effectively, resulting in captions that closely resemble human descriptions.

This model also utilizes teacher forcing during training to accelerate convergence and improve language fluency. Standard evaluation metrics, such as BLEU, are employed to quantitatively assess performance and ensure comparability with existing models. Models inspired by works such as *"Show, Attend and Tell"* highlight the importance of attention in bridging the gap between image features and natural language generation.

Attention-based captioning systems have the potential to serve a wide range of real-world applications, including assistive tools for the visually impaired, intelligent content indexing, and human–computer interaction. With the growing interest in multimodal learning and transformer-based architectures, future research in this area is likely to focus on scaling these models to larger datasets and improving generalization across diverse visual scenes.

Acknowledgment

We thank everyone who helped us in various means and stood up as a support system whenever needed.

References

[1] Y. Ming *et al.*, "Visuals to text: A comprehensive review on automatic image captioning," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 8, pp. 1339–1365, Aug. 2022, doi: 10.1109/JAS.2022.105734.

[2] A. Gupta, D. S. Bhadauria, M. Atray, and I. Kaur, "Predicting relevant captions using image caption generator in social media platforms," in *Computational Methods in Science and Technology*. Boca Raton, FL, USA: CRC Press, 2024, pp. 423–432, doi:10.1201/9781003501244-65.

[3] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent neural networks: A comprehensive review of architectures, variants, and applications," Information, vol. 15, no. 9, p. 517, 2024, doi: 10.3390/info15090517.

[4] S. Pandey, P. Saha, and G. Sharan, "Enhancing chest X-ray analysis using encoder-decoder with GRU for report generation," in *Proc. 4th Int. Conf. Adv. Electr., Comput., Commun. Sustain. Technol. (ICAECT)*, Bhilai, India, 2024, pp. 1–8, doi:10.1109/ICAECT60202.2024.10469644.

[5] H. Suh, J. Kim, J. So, and J. Jung, "A core region captioning framework for automatic video understanding in story video contents," *Int. J. Eng., Bus. Manag.*, vol. 14, Nov. 2022, doi:10.1177/18479790221078130.

[6] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 740–755, doi: 10.1007/978-3-319-10602-1_48.

[7] Adityajn "Flickr8K Dataset," Kaggle. Accessed: [Date]. [Online]. Available: https://www.kaggle.com/datasets/adityajn105/flickr8k.

[8] O. Arshi and P. Dadure, "A comprehensive review of image caption generation," *Multimed. Tools Appl.*, vol. 84, no. 25, pp. 29419–29471, 2025, doi: 10.1007/s11042-024-20095-0.

[9] M. Mahajan *et al.*, "Image captioning—A comprehensive encoder-decoder approach on Flickr8K," in *Proc. Int. Conf. Autom. Comput. (AUTOCOM)*, Dehradun, India, 2025, pp. 1310–1315, doi:10.1109/AUTOCOM64127.2025.10956672.

[10] K. R. Suresh, A. Jarapala, and P. V. Sudeep, "Image captioning encoder-decoder models using CNN-RNN architectures: A comparative study," *Circuits, Syst., Signal Process.*, vol. 41, no. 10, pp. 5719–5742, Oct. 2022, doi: 10.1007/s00034-022-02050-2.

[11] A. Raphael, S. Abisri, E. Anitha, S. Ritika, and M. Venugopalan, "Attention based CNN-RNN hybrid model for image captioning," in *Proc. IEEE 5th Glob. Conf. Adv. Technol. (GCAT)*, Bangalore, India, 2024, pp. 1–5, doi: 10.1109/GCAT62922.2024.10923871.

[12] H. Parmar, M. Rai, and U. K. Murari, "A novel image caption generation based on CNN and RNN," in *Proc. Int. Conf. Adv. Comput. Res. Sci., Eng. Technol. (ACROSET)*, Kozhikode, India, 2024, pp. 1–8, doi: 10.1109/ACROSET62108.2024.10743848.

[13] T. Zhang, T. Zhang, Y. Zhuo, and F. Ma, "CATNIC: A feature relevance based transformer model for automatic image caption generation," *SSRN Electron. J.*, 2022, doi: 10.2139/ssrn.4272712.

[14] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Boston, MA, USA, Jun. 2015, pp. 3156–3164, doi: 10.1109/CVPR.2015.7298935.

[15] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, "End-to-end attention-based large vocabulary speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Shanghai, China, Mar. 2016, pp. 4945–4949, doi:10.1109/ICASSP.2016.7472618.

[16] M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion RNN-LSTM architecture for image caption generation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Phoenix, AZ, USA, Sep. 2016, pp. 4448–4452, doi: 10.1109/ICIP.2016.7533201.

[17] S. Sheng and M.-F. Moens, "Generating captions for images of ancient artworks," in *Proc. 27th ACM Int. Conf. Multimedia*, Nice, France, Oct. 2019, pp. 2478–2486, doi: 10.1145/3343031.3350972.

[18] J. R. Chowdhury and C. Caragea, "Beam tree recursive cells," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Honolulu, HI, USA, Jul. 2023, pp. 28768–28791.

[19] S. Kudugunta *et al.*, "Madlad-400: A multilingual and document-level large audited dataset," *Adv. Neural Inf. Process. Syst.*, vol. 36, pp. 67284–67296, Dec. 2023.

[20] S. Liu and J. Zhang, "Local alignment deep network for infrared-visible cross-modal person reidentification in 6G-enabled Internet of Things," *IEEE Internet Things J.*, vol. 8, no. 20, pp. 15170–15179, Oct. 2021, doi: 10.1109/JIOT.2020.3038794.

[21] N. K. Kumar, D. Vigneswari, A. Mohan, K. Laxman, and J. Yuvaraj, "Detection and recognition of objects in image caption generator system: A deep learning approach," in *Proc. 5th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, Coimbatore, India, Mar. 2019, pp. 107–109, doi: 10.1109/ICACCS.2019.8728516.