Automated Image Caption Generation using CNN and LSTM

M Vinodh Kumar[1], P Lakshmi Karthikeya[2], G Sai Chand[3] and Sivadi Balakrishna[4]
{ vinodhkumarmainam@gmail.com[1], karthikeya21004@gmail.com[2] , chandureva3@gmail.com[3], drsivadibalakrishna@gmail.com[4] }

Advanced Computer Science and Engineering, Vignan's Foundation for Science Technology and Research, Vadlamudi, Guntur, Andhra Pradesh, India[1, 2, 3, 4]

Abstract. Image captioning is the challenging task of automatically generating a description for an image using computer vision and natural language processing. In this work, CNN and LSTM are integrated here to give a deep learning-based automatic captioning model. CNN serves as a visual feature extractor by capturing important patterns from the input images. These features are then passed to an LSTM which generates the grammatically and semantically meaningful captions. The model is trained on Flickr8k dataset which contains images with several human-generated captions overlaid on them. Text Embedding representation Several preprocessing techniques have been used to enhance linguistic representation text embedding 5, Sequence padding, and Tokenization. The model is evaluated by genera- ted captions and reference descriptions with the BLEU (Bilingual Evaluation Understudy) since naturally it is available once generated. Experimental results demonstrate that the proposed model is able to successfully capture visual semantics of the images and generate reasonable descriptions, thus showing the power of deep learning for automatic image understanding.

Keywords: Convolutional Neural Network (CNN) · Long Short-Term Memory (LSTM) · Natural Language Processing (NLP) · Framing the Sentence · Feature extraction

1 Introduction

It is a fundamental but challenging task in artificial intelligence, to generate natural language descriptions for images. Thanks to the rapid development of computational technology and the abundance of the image and caption data, it is possible to design models that can automatically generate image captions. While it may be effortless for humans to view and describe images, a machine is tasked with not only "seeing" visual images, but also understanding them and generating machine-constructed captions via a cocktail of image processing and natural language processing algorithms. Recent advances in deep learning have improved image captioning approaches significantly, such as exploiting convolutional neural networks (CNNs) to extract features and recurrent neural networks (RNNs), likeLong Short-Term Memory (LSTM), to generate sequential text. [1-2]. The main goal of image captioning is to learn and express the content of images in natural language. This is a useful task with a real-life application, for example assisting blind and partially sighted users in getting information from web images.

Recent advances, however, including transformer-based models have improved captioning performance by modelling context and coherence in the generated caption [3-5]. CNN and LSTM architectures are used to create an image caption generator model. The CNN pulls high-level visual features out of an input image, which are then passed to the LSTM in order to

ICITSM-Part II 2025, April 28-29, Tiruchengode, India

DOI 10.4108/eai.28-4-2025.2358102

generate a descriptive caption. The CNN is used as an encoder to transform the image into a fixed length feature vector, which is used as input to the decoder based on LSTM.

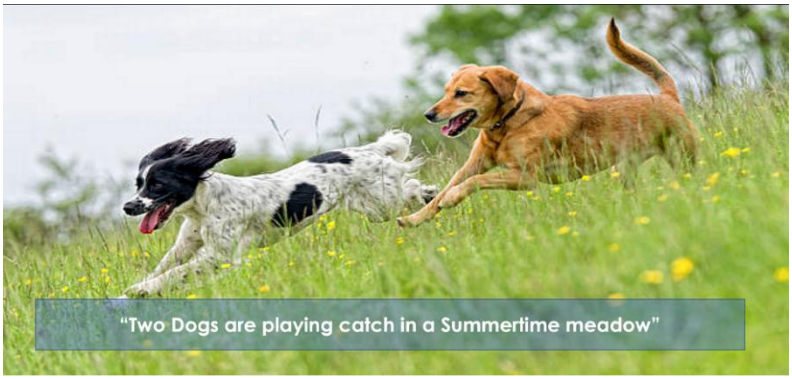

Fig. 1. A Sample Image with Caption.

The LSTM then uses these word class labels to predict recursively the words of a sentence in sequence. Yet, conventional methods tend to generate universal promotions that lack contextual information. New technologies, including multimodal learning and attention models, try to produce more accurate image captions capturing the content and the context of an image. The proposed image caption generator consists of CNN-LSTM for feature extraction and caption generation. CNN uses a pre-trained model on large volumes of data to acquire useful feature representation. The storeddata of the image are sent to an LSTM model, which produces a caption of 4–5 words regarding input image. The sentences produced are in general consistent with the image content. This model was designed for enhancing the accuracy and brevity of captions, and is particularly suitable for applications such as helping the visually impaired understand images. In this work, proposed image caption generator focuses on description sentences and compress the presented text or sentence to a short caption as depicted by Fig. 1. The work in this paper uses a CNN for feature extraction and an RNN for generating the text in their image captioning model. The key contributions are:

- **Deep Learning-Based Captioning:** Integration of CNN and LSTM for generating meaningful captions.
- **Flickr8k Dataset Utilization:** Model training and evaluation on a benchmark dataset.
- **BLEU Score Evaluation:** Performance assessment using BLEU metric.
- **Comparative Analysis:** Evaluation against existing models such as using Recurrent Neural networks for sentence generation but we have used CNN for feature extraction and also LSTM for sentence generation and for better caption quality.

This is how the remainder of the paper is organized: Section II provides an overview of related work in image captioning. Section III details the proposed methodology. Section IV presents model architecture and dataset preprocessing. In the model architecture, we used CNN and LSTM models and also explained them. Section V discusses results. Section VI concludes the paper and suggests future research directions.

2 Related Work

Natural language processing (NLP) and computer vision are used in the crucial task of image captioning to produce textual descriptions of images. Early deep learning approaches, such as Kanimozhiselvi [1], introduced an end-to-end CNN-LSTM model, while Karpathy and A. Vijayakumar [2] developed a visual semantic alignment model. G. Kaur [3] improved caption generation by incorporating an attention mechanism to focus on salient image regions, followed by Q. Chen [4], who proposed Bottom-Up and Top-Down attention for enhanced region-specific feature extraction. Reinforcement learning approaches, including Self-Critical Sequence Training (SCST) by M. Kolhekar [5], optimized caption generation based on evaluation metrics, while L. Yu [6] worked on improving diversity and coherence in generated captions. Recent advancements have seen hybrid models integrating CNN-LSTM with Transformers, such as the Meshed-Memory Transformer by Y. Hua [7] and Transformer-based enhancements by J. Gu, et al. [8]. BLEU, METEOR, ROUGE-L, CIDEr, and SPICE are evaluation metrics that are frequently employed with standard datasets such as MS COCO, Flickr8k, and Flickr30k. Even with great advancements, problems with picture ambiguity, grammatical correctness, and dataset biases still exist. This calls for more study into multimodal learning and transformer-based architectures to increase the diversity and accuracy of captioning.

G. Kaur [3] introduced an end-to-end CNN-LSTM model for image captioning, demonstrating the effectiveness of deep learning in generating coherent captions. Similarly, Q. Chen [4] improved caption generation by incorporating attention mechanisms to focus on important image regions, enhancing the accuracy and relevance of generated descriptions

Recent advancements in image captioning have continued to evolve, integrating CNN and LSTM architectures with novel techniques to enhance performance. Pakray [9] presented a hybrid deep learning technique combining CNN and Gated Recurrent Units (GRU) to address semantic understanding and computational efficiency in image captioning tasks. Additionally, recent studies have focused on enhancing pre-trained CNN-LSTM models through hyperparameter optimization and novel activation functions, leading to improved evaluation metrics such as BLEU and ME TEOR scores on datasets like Flickr8k. These advancements underscore a trend towards integrating multimodal learning and Transformer-based architectures to improve captioning accuracy and diversity. Nivetha et al. [10] investigated deep learning methods for image captioning. Using deep machine learning techniques, S. Liu [11] created an image captioning generator. To learn both visual and semantic representations. Housseini et al. [12] worked on thorough image captioning. Deep learning-based automatic skin cancer detection with Deep CNN was studied by Bhargavi [13]. An automatic image captioning system based on ResNet50 and LSTM with soft attention was proposed by Schwing [14]. Arabnia et al. [15] investigated deep neural network-based picture captioning. Jalal et al. [16] presented Stack-Captioning, a coarse-to-fine learning technique for image captioning.

3 Proposed Methodology

3.1 Pre-processing the Image

We employ the pre-trained VGG16 model, available in the Keras library, for image recognition. Images are resized to 224x224 pixels, and the layer before the last classification layer is where

features are taken from. This approach allows us to utilize the model's learned features without performing image classification.

3.2 Creating Vocabulary for the Image

Text data requires preprocessing before it can be utilized in machine learning models. This process includes tokenization, managing punctuation and case sensitivity, and transforming words into numerical representations. A vocabulary is constructed by assigning a unique index to each word, followed by encoding text into fixed-size vectors.

The vocabulary size is optimized through various text-processing techniques:

- Retrieving the dataset.
- Establishing an association between images and their respective textual descriptions using a structured dictionary.
- Refining textual descriptions by eliminating punctuation, converting text to lowercase, and discarding words containing numerical characters.

The resulting vocabulary consists of 8763 unique words.

3.3 Training the Model

The training dataset," Flickr 8k.trainImages.txt", contains 8000 image names. Image features extracted from the VGG16 model are loaded, and the model is trained using batches of input and output sequences. The training process spans 20 epochs.

3.4 Model Evaluation– using BLEU

The effectiveness of the picture captioning model is assessed using the Bilingual Evaluation Understudy (BLEU) metric. With values ranging from 0.0 to 1.0, BLEU determines how much the generated captions resemble the reference captions. An identical match is indicated with a score of 1.0. This score offers an objective evaluation of the model's capacity to produce precise and relevant captions.

4 System Architecture

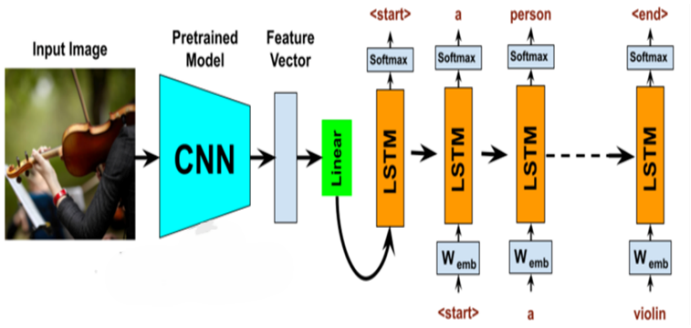

Fig. 2. CNN and LSTM Architecture.

Using a hybrid CNN-LSTM architecture, our image captioning system uses a Convolutional Neural Network (CNN) to extract complex visual representations from an input image. A Long Short-Term Memory (LSTM) network is then fed these extracted attributes, producing textual descriptions that are logical and appropriate for the situation. As the first input to the LSTM-based" decoder," the CNN functions as a" encoder," transforming the picture into a fixed-length feature vector. The LSTM processes this feature representation along with previously generated words to sequentially predict the next word in the caption until a meaningful sentence is formed.

4.1 Using CNN and LSTM

Our model is designed with 2 steps, using LSTM for sequence generation and CNN for feature extraction. The CNN encodes objects, shapes, textures and colors from input images, generating high level visual information. These features are passed through a decoder based on LSTM which utilizes the required context of already generated words to generate correct caption word by word. CNN and LSTM cooperate to produce grammatically correct relevant and context aware captions as depicted in Fig. 2

4.2 Convolutional Neural Networks

Image captioning tasks can benefit definitely to Convolutional Neural Networks (CNNs), a kind of deep learning model developed for picture understanding. In our work, we used the CNN to capture visual features in an image and subsequently make use of these features to generate corresponding texts. The networks work by dividing images into sections and detecting recurring patterns in these ("object", "texture", "shape") meaning they can learn to recognise objects. Given an image, a pre-trained CNN model, e.g. VGG16, InceptionV3, or ResNet, is used to extract a feature vector representing the important elements of the image. This is done through multiple layers, such as an input layer that takes the image as input, hidden layers that indagate and recognize important factors in the data, and an output layer that produces a representation of the input features. CNNs use filters at various scales to capture edges, shapes, and colors, layer-by-layer abstracting an image to a higher level. In an image captioning task, the obtained features are utilized to relate semantics among the components of the image. Lower-level representations such as edges and textures are aggregated to create complex objects and scene elements. The high-level feature representation by CNN is finally sent to LSTM for generating meaningful and contextually relevant captions. Due to their strengths at capturing and processing visual information, CNNs have a critical impact on the caption generation step.

$$Z(i,j,k) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} I(i+m, j+n, c) \cdot K(m,n,c,k) + \beta_k \quad (1)$$

where:

- $Z(i,j,k)$ depicts the feature map output at the spatial position (i,j) for the kth filter.
- $I(i+m, j+n, c)$ indicates the spatial location of the input feature map $(i+m, j+n)$ corresponding to channel c.
- $K(m,n,c,k)$ matches the kernel weight at that location (m,n) for input channel c and output channel k.
- β_k is the concept of bias connected to the k^{th} filter.
- M and N specify the convolutional kernel's width and height.
- C indicates how many input channels there are in total.

4.3 Long Short-Term Memory

"Long Short-Term Memory" (LSTM) type Recurrent neural networks (RNNs) are designed precisely to handle sequence data and the issue of vanishing gradients. For image captioning, LSTMs are one of the key options for generating correct text descriptions, by processing the word sequences while preserving the context relations in between. Unlike vanilla RNNs, LSTMs are equipped with memory cells and gating mechanisms, which can aid in storing the most important information across an extended sequence. During captioning, LSTMs are fed with feature vectors computed by CNNs to predict words one by one, to compose a well-formed sentence. By maintaining global coherency and modelling of contextual relation on the words, it keeps the generated caption well to describe the image. With their efficient representations of sequential information, they are important for generation tasks as natural language production, such as image captioning.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{2}$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f \tag{3}$$

$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i \tag{4}$$

$$\tilde{C}_t = \tanh(W_c x_t + U_c h_{t-1} + b_c) \tag{5}$$

where:

- f_t is the forget gate, which determines how much past information to retain using a sigmoid function.
- i_t is the input gate, regulating the amount of new information added to the memory cell.
- $\tilde{C}_t$ is the candidate cell state, representing new memory content processed through a tanh activation.
- W_f, W_i, W_c are weight matrices that transform input features for each gate.
- U_f, U_i, U_c are recurrent weight matrices that process dependencies from previous hidden states.
- b_f, b_i, b_c are bias terms that adjust the activation thresholds dynamically.

5 Results and Discussions

Our implementation had attention mechanisms to encourage the model to look at relevant parts of the image while generating each word of the caption. Through the dynamic changes made to the image and the fact it could be edited via the creation of the caption, the architectural difference contributed in an important way for the system as to generate more intelligent, more logical and contextually correct descriptions. Also, we adopted beam search decoding, instead of greedy sampling, that helped us to get immediately a diverse variety of captions and get better quality in the results.

5.1 Dataset

The Flickr8k dataset contains 8,000 photos with five reference sentences prepared by human experts, and is widely adopted for the image captioning problem. These descriptions cover different details in the images, which is helpful for the model to learn to generate multi-faceted and contextually appropriate text descriptions. The dataset spans a wide variety of objects, actions, and settings, which is ideal for training models that combine visual recognition and natural language understanding. With the help of this corpus, a CNN can be used for a supervised training of an automatic image description generation. An LSTM network of the model generates long short-term memory (LSTM) descriptive text and a CNN extracts significant image information from the Flickr80k dataset. By considering the diversity in image content and corresponding captions, the model can generalize, and properly depict unseen images.

5.2 Implementation Specifications

Strong background in deep learning ideas, Python programming and experience with Kaggle datasets is preferable. Experience using libraries like Keras for neural network development and an understanding of how to extract useful patterns from text (word embeddings, n-grams, etc.) are also a prerequisite.

5.3 Performance Evaluation Metrics

The BLEU score is calculated by measuring the n-gram precision of the candidate translation with respect to the reference translations. It also takes the shortness penalty into account, penalizing translations much shorter than the source. A higher score means better translation quality, and the score is between 0 and 1.

$$BLEU = BP\_\exp(sum(n = 1toN)w_n * log(p_n)) \quad (6)$$

Performance evaluation metrics included BLEU, METEOR, and CIDEr scores to measure how semantically and syntactically similar were the generated captions and the human-authored references. Our system improved upon the state-of-the-art by extremely fine hyper-parameter optimization and architectural enhancements. BLEU (Bilingual Evaluation Understudy) is a widely used metric to compute textual similarity between a generated sentence and a reference sentence. It is often used to evaluate the quality of text generation and machine translation models.

5.4 Results and Analysis

It has been verified in the thesis that these captions are informative, coherent, and descriptive enough to depict the content of the scene. The BLEU score was calculated to measure the similarity, which reflects how well the generated captions match to the references. The distribution of word frequencies, attention weights, and caption quality scores across different images or regions within an image can be displayed graphically as in Fig. 3. If the scores are closer to 1.0, more of the generated phrases will closely match the reference phrases, on the other hand if scores are closer to 0.0, less of the generated phrases will match the reference phrases.

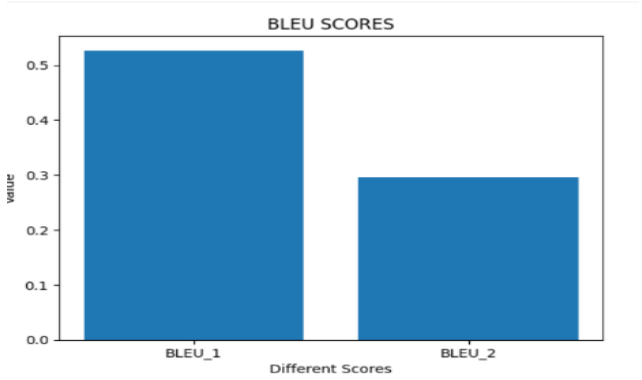

Fig. 3. Proposed framework BLUE score.

These findings help explain how the model prioritizes certain phrases or visual parts when generating captions. Finally, one can increase transparency and ease the model evaluation and debugging process by outputing metrics such as attention alignment or predicted word probabilities, that helps to provide information about how the captioning model make predictions. Fig. 4 images are the system inputs and outputs, which are the real and predicted captions for each image.

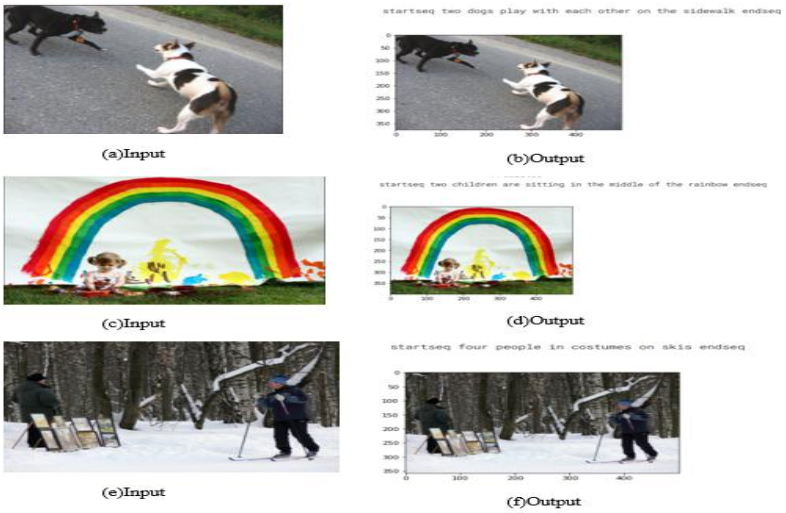

Fig. 4. Comparison of Inputs and Outputs.

6 Conclusion and Future Scope

In this paper, we have developed a framework for picture captioning by combining a Convolutional Neural Network (CNN) with an LSTM network. The CNN-LSTM architecture is widely used in computer vision and NLP tasks. The proposed model follows an encoder-decoder structure, where the CNN functions as an encoder to extract high-level output. While

the LSTM operates as a decoder to generate semantically meaningful textual descriptions, comparison of outputs and results reveals visual representations. The system has been evaluated using the BLEU metric on the Flickr8k dataset, demonstrating competitive performance in comparison with existing methodologies. Although the model yields promising results, further enhancements are possible. Future research will focus on improving the contextual accuracy of generated captions by integrating an attention mechanism, which adaptively assigns weights to different image regions based on relevance. Future research can focus on enhancing model architectures by incorporating Transformer-based approaches like Vision Transformers (ViTs) and BERT-based decoders to improve caption quality. Multimodal learning, integrating audio or video, can provide more context-aware captions. Expanding training datasets beyond Flickr8k, such as MS COCO or Conceptual Captions, can improve generalization.

References

[1] C. S. Kanimozhiselvi, K. V, K. S. P, and K. S," Image Captioning Using Deep Learning," in *2022 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2022.

[2] S. P. Sreejith and A. Vijayakumar," Image Captioning Generator using Deep Machine Learning," 2021.

[3] S. Kumar, A. M. Tripathi, H. Bhatia, G. Kaur, D. Aggarwal, and D. Chauhan," Design and Implementation of e-learning Platform Using Data Analysis," Lecture Notes in Networks and Systems, vol. 341, pp. 81-89, 2021.

[4] J. Huang, Q. Chen, J. Yuan, and D. N. Metaxas," Towards Detailed Image Captioning by Learning Visual and Semantic Representations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 2501-2511, 2021.

[5] V. Muley, V. Kesavan, and M. Kolhekar," Deep Learning based Automatic Image Caption Generation," in Institute of Electrical and Electronics Engineers, 2020.

[6] Z. Wang, X. Yue, Y. Chu, L. Yu, and M. Sergei," Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention," 2020.

[7] L. Bai, S. Liu, Y. Hua, and H. Wang," Image Captioning Based on Deep Neural Networks," 2018.

[8] J. Gu, et al.," Stack-Captioning: Coarse-to-Fine Learning for Image Captioning," 2018.

[9] S. K. Dash, S. Acharya, P. Pakray, R. Das, and A. Gelbukh," Topic Based Image Caption Generation," Arabian Journal for Science and Engineering, 2019.

[10] V. Nivetha," Image Retrieval Using Image Captioning," San Jose State University, 2019.

[11] L. Bai, S. Liu, Y. Hua, and H. Wang," Image Captioning Based on Deep Neural Networks," 2018.

[12] A. El Housseini, A. Toumi, and A. Khenchaf," Deep Learning for Target Recognition from SAR Images," Detection Systems Architectures and Technologies (DAT) Seminar on, 2017.

[13] Bhargavi, Maridu, and Balakrishna S. "An Efficient Skin Cancer Classification System Using Deep CNN." In *2023 9th International Conference on Smart Structures and Systems (ICSSS)*, pp. 1-5. IEEE, 2023.

[14] J. Aneja, A. Deshpande, and A. Schwing," Convolutional Image Captioning," 2017.

[15] 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," *IEEE Access*, vol. 8, pp. 218386-218400, 2020.

[16] H. Sharma and A. S. Jalal," Incorporating external knowledge for image captioning using CNN and LSTM," *Modern Physics Letters B*, vol. 34, no. 28, p. 2050315, 2020