WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Deep Learning-Based Image Captioning: A Hybrid CNN-LSTM Approach.

V. Pravallika[1], V. Uday Kiran[2], B. Rahul[3], N. Neelima[4], G. Rishi Patnaik[5], DR. Sreejyothshna Ankam[6]

[1,2,3,4,5] UG Student, Department of CSE(AI &ML),
[6]Senior Assistant Professor, Department of CSE(AI &ML)

ABSTRACT

In today's digital age, image captioning has become a crucial tool for bridging the gap between visual content and natural language. This project aims to develop an image caption generation model that automatically produces descriptive and coherent text for a given image. Image captioning plays a significant role in computer vision and natural language processing, with applications in platforms like Facebook and Google Photos for image segmentation and organization. Additionally, it can automate tasks that require human interpretation of images, making it valuable for accessibility and content management. To achieve this, the project utilizes deep learning techniques, specifically Convolutional Neural Networks (CNN) for visual feature extraction and Long Short-Term Memory (LSTM) networks for sequential text generation. The model is trained on the Flickr8k dataset, which contains 8,000 images, each paired with five captions. This approach enhances the accuracy and relevance of generated captions, making it useful for various real-world applications in accessibility, search engines, and multimedia platforms.

KEYWORDS: Image Captioning, Convolutional neural network (CNN), Long Short-Term Memory (LSTM), Deep learning, Flickr8k Dataset.

INTRODUCTION

Image caption generation is a fundamental task at the intersection of computer vision and natural language processing, aimed at automatically generating descriptive textual information from visual content. This task requires a model to not only recognize objects, scenes, and actions within an image but also to construct coherent, contextually relevant sentences that accurately describe the visual elements. The ability to generate meaningful captions has significant applications in various domains, including assisting visually impaired individuals, improving image retrieval systems, enhancing human-computer interaction, and facilitating content indexing.

One of the primary challenges in image captioning is bridging the gap between vision and language. Understanding an image involves extracting relevant features while also interpreting the spatial relationships between different objects. Generating a caption further requires forming grammatically correct and semantically meaningful sentences. Recent advancements in deep learning have enabled effective solutions by integrating Convolutional Neural Networks (CNNs) for feature extraction with Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, for sequential text generation. CNNs excel at capturing spatial hierarchies in images, making them highly suitable for visual representation learning, while LSTMs effectively model sequential dependencies, making them ideal for generating coherent captions.

In this work, VGG16, a widely used CNN architecture, is utilized to extract meaningful visual features from images. These extracted features are then processed by an LSTM network to generate corresponding textual descriptions. The Flickr8k dataset, consisting of 8,000 images, each annotated with five diverse captions, serves as the benchmark for training and evaluation. The presence of multiple captions per image allows the model to learn various linguistic expressions, enhancing its ability to produce accurate and fluent descriptions.

The primary objective of this research is to develop an image captioning model capable of generating high-quality, contextually rich, and semantically meaningful captions. By leveraging deep learning techniques, improvements in accuracy, diversity, and fluency of the generated captions are targeted. This research contributes to the growing field of automated image captioning and paves the way for more advanced multi-modal AI applications that seamlessly integrate vision and language.

LITERATURE SURVEY

Panicker et al. explored image captioning using CNNs for feature extraction and LSTMs for text generation, employing the Flickr8k dataset. They utilized an encoder-decoder framework to improve contextual understanding and evaluated models using BLEU scores [1].

Keshamoni and Priyanka discussed the evolution of image captioning from rule-based systems to deep learning models, emphasizing applications in assisting the visually impaired and medical image analysis. They highlighted the role of transfer learning with models like VGG-16 and ResNet in improving caption quality [2].

Mishra et al. investigated the use of Vision Transformers and GPT architectures for image captioning, replacing traditional CNN-LSTM models. They emphasized the role of attention mechanisms and multimodal pre-training with models like BERT and TAP [3].

Verma et al. reviewed deep learning approaches in image captioning, comparing template-based, retrieval-based, and deep learning methods. They highlighted advancements in models like NIC and m-RNN, which achieved high BLEU scores [4].

Yoga et al. integrated CNN-LSTM models with linguistic analysis to enhance caption generation. They evaluated architectures like ResNet-LSTM and emphasized the importance of attention mechanisms for focusing on relevant image regions [5].

Eagalapati et al. proposed integrating image captioning into ChatGPT to improve its understanding of visual content. Their work introduced agent-centric approaches using natural language quantifiers [6].

Patil et al. analyzed the performance of CNN-LSTM models for image caption generation, emphasizing end-to-end training. They referenced studies by Swarnim and Ravi (2021) and Sreejith and Vijayakumar (2021) on the effectiveness of CNN-LSTM approaches [7].

Raypurkar et al. demonstrated the effectiveness of CNNs and LSTMs in image captioning, using datasets like MS COCO. They highlighted advancements in visual attention mechanisms for improving context in captions [8].

Chohan et al. systematically reviewed deep learning techniques in image captioning, emphasizing encoder-decoder architectures and attention mechanisms. They discussed the use of datasets like MS COCO, Flickr8k, and Flickr30k. Evaluation metrics such as BLEU, ROUGE-L, and CIDEr were highlighted [9].

Sharma et al. explored the shift from template-based methods to RNNs and LSTMs for image captioning. They addressed challenges like the vanishing gradient problem and compared inject and merge architectures. The merge architecture was preferred for its simplicity and efficiency [10].

Kesavan et al. investigated the use of CNNs and RNNs in the "Show and Tell" model for image captioning. They introduced attention mechanisms in the "Show, Attend and Tell" model to focus on specific image parts. The need for large, well-annotated datasets was emphasized [11].

Amritkar and Jabade categorized image captioning methods into template-based, transfer-based, and neural network approaches. They highlighted the effectiveness of multimodal RNNs like the NIC model and the integration of visual attention in LSTMs [12].

Ding et al. proposed a multi-modal neural language model combining CNNs and NLP for image captioning. They discussed LSTM-based architectures for large-scale visual learning and tree-structured methods for generating descriptions [13].

Oluborode et al. explored deep learning frameworks for image captioning using CNN-LSTM architectures and datasets like Flickr8K. They evaluated models using metrics like BLEU and METEOR. Hybrid models combining CNNs and RNNs were noted to enhance caption generation [14].

Padate et al. introduced dual attention mechanisms combining CNN and BI-LSTM for image captioning. They used the SI-EFO algorithm to optimize BI-LSTM weights, improving performance over existing methods [15].

RESEARCH METHODOLOGY

VGG16

VGG16 is a deep convolutional neural network (CNN) architecture introduced by the Visual Geometry Group at Oxford. It consists of 16 layers, including 13 convolutional layers with small 3×3 filters, followed by max pooling layers and three fully connected layers. The design emphasizes deep feature extraction while maintaining a uniform architecture. Trained on the ImageNet dataset, it achieves high accuracy in image classification and is widely used for transfer learning in various computer vision applications, such as object detection, image captioning, and feature extraction.

LSTM

Long Short-Term Memory (LSTM) is a type of Recurrent Neural Network (RNN) designed to handle sequential data while overcoming the vanishing gradient problem in standard RNNs. It consists of memory cells with gates (input, forget, and output) that regulate the flow of information, allowing it to retain long-term dependencies. LSTMs are widely used in natural language processing (NLP), speech recognition, time-series forecasting, and image captioning, where sequential patterns are crucial. Their ability to remember long-term dependencies makes them highly effective for tasks requiring contextual understanding.

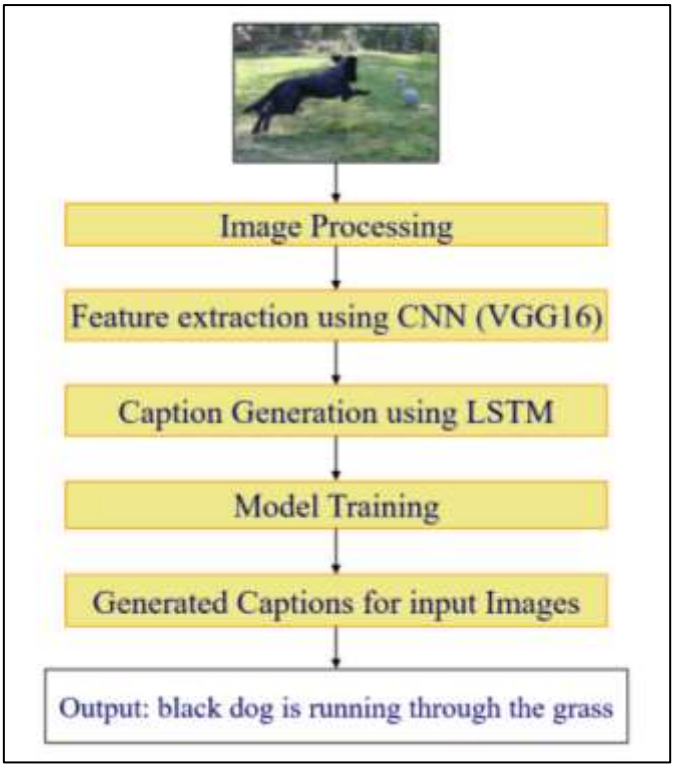

Fig: Work flow of image caption generator

The development of the image captioning model follows a structured methodology that integrates deep learning techniques for both visual feature extraction and text generation. The process begins with dataset collection and preprocessing, where the Flickr8k dataset is utilized, consisting of 8,000 images, each annotated with multiple descriptive captions. To prepare the data for training, images are resized to a fixed dimension and their pixel values are normalized to enhance model efficiency. Caption preprocessing involves converting text to lowercase, removing punctuation, tokenizing words, and mapping words to a vocabulary to ensure uniformity and facilitate sequence modeling.

Once the data is preprocessed, feature extraction using a Convolutional Neural Network (CNN) is performed. A pre-trained VGG16 model, known for its strong image classification capabilities, is used to extract high-level visual features from images. The fully connected layers of VGG16 are removed, retaining only the convolutional layers to capture essential spatial features. These extracted feature vectors serve as inputs to the caption generation model.

For caption generation, a Long Short-Term Memory (LSTM)-based Recurrent Neural Network (RNN) is employed to generate meaningful text descriptions based on the extracted image features. The model incorporates a word embedding layer, which transforms words into dense vector representations, allowing the LSTM to understand semantic relationships between words. The LSTM sequentially predicts words, generating captions in a structured manner while maintaining contextual coherence.

During model training, both the CNN and LSTM components are trained together using categorical cross-entropy loss, which helps minimize prediction errors. The Adam optimizer is used for efficient gradient updates, ensuring faster convergence. The model undergoes training over multiple epochs, allowing it to refine its ability to generate accurate and contextually rich captions. To prevent overfitting, techniques such as dropout regularization and early stopping may be applied.

Finally, in the caption generation and evaluation phase, test images are passed through the trained model to generate captions. The generated captions are evaluated using natural language processing (NLP) metrics, including BLEU (Bilingual Evaluation Understudy) scores, which measure the similarity between generated and human-annotated captions based on n-gram precision. A higher BLEU score indicates better caption quality. The results are analyzed to assess the model's effectiveness in generating diverse, fluent, and contextually accurate image descriptions. This structured methodology ensures that the image captioning model is both robust and efficient in understanding visual content and generating meaningful textual descriptions.

RESULTS

The proposed hybrid CNN-LSTM model for image caption generation was evaluated on the Flickr8k dataset, demonstrating its ability to produce descriptive and contextually relevant captions. The model's performance was assessed using the BLEU (Bilingual Evaluation Understudy) metric, which measures the similarity between the generated captions and the reference captions provided in the dataset. Below is a comparison table with three examples of original and predicted captions, showcasing the model's ability to generate meaningful descriptions for diverse images. These results highlight the effectiveness of the CNN-LSTM approach in bridging the gap between visual content and natural language, making it a promising tool for applications in accessibility, image retrieval, and multimedia platforms. Further enhancements, such as incorporating attention mechanisms or leveraging larger datasets, could improve the model's performance in handling finer details and more complex scenes.

Table: Comparison between original and predicted captions

Image	Original caption	Predicted caption
	Man in hat is displaying pictures next to skier in blue hat.	Man on skis looking at framed pictures set up in the snow.
	Boy in blue swimming trunks slides down yellow slide into wading pool with inflatable toys floating in the water.	Boy sliding down slide into pool with colorful tubes.
	Little girl is sitting in front of large painted rainbow.	Little girl is sitting in front of large painted rainbow.

CONCLUSION

This research presented a hybrid CNN-LSTM model for automated image caption generation, combining Convolutional Neural Networks (CNNs) for visual feature extraction and Long Short-Term Memory (LSTM) networks for sequential text generation. Trained and evaluated on the Flickr8k dataset, the model demonstrated its ability to generate grammatically correct and contextually relevant captions. While the model effectively identifies key objects and their relationships, there is potential for improvement in capturing finer details and complex sentence structures. The results highlight the applicability of this approach in real-world scenarios such as accessibility, image retrieval, and multimedia content management. Future work could explore advanced techniques like attention mechanisms or transformer-based architectures to further enhance performance, contributing to the development of more sophisticated multi-modal AI systems that bridge vision and language.

REFERENCES

[1] Panicker, M. J., Upadhayay, V., Sethi, G., & Mathur, V. (2021). Image caption generator. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *10*(3), 87-92.

[2] Keshamoni, K., & Priyanka, P. (2024, March). Visionary Narratives: Unveiling the Potential of an Automated Image Caption Generator through Deep Learning and Multidimensional Analysis. In *2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI)* (Vol. 2, pp. 1-6). IEEE.

[3] Mishra, S., Seth, S., Jain, S., Pant, V., Parikh, J., Jain, R., & Islam, S. M. (2024, May). Image Caption Generation using Vision Transformer and GPT Architecture. In *2024 2nd International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 1-6). IEEE.

[4] Verma, A., Yadav, A. K., Kumar, M., & Yadav, D. (2024). Automatic image caption generation using deep learning. *Multimedia Tools and Applications*, *83*(2), 5309-5325.

[5] Yoga, M., Ramyasri, M. M., Bhavatharini, N., Harini, M., & NivethaSri, A. (2024, April). Unravelling Visual Narratives with Deep Learning for Image Caption Generation. In *2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS)* (pp. 387-393). IEEE.

[6] Eagalapati, R. C. K., Chowdary, M. C., Sasank, A. S., Monish, B., & Kumar, P. R. (2024, April). Enriching Conversations: Empowering ChatGPT with Image Caption Generation. In *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)* (pp. 1-5). IEEE.

[7] Patil, S. S., Varma, B. S., Devadasu, G., Basha, C. H., Inamdar, M. J. R., & Salman, S. S. (2022, August). Performance analysis of image caption generation using deep learning techniques. In International Conference on Microelectronic Devices, Circuits and Systems (pp. 159-170). Cham: Springer Nature Switzerland.

[8] Raypurkar, M., Supe, A., Bhumkar, P., Borse, P., & Sayyad, S. (2021). Deep learning based image caption generator. International Research Journal of Engineering and Technology (IRJET), 8(03).

[9] Chohan, M., Khan, A., Mahar, M. S., Hassan, S., Ghafoor, A., & Khan, M. (2020). Image captioning using deep learning: A systematic. image, 11(5), 62.

[10] Sharma, G., Kalena, P., Malde, N., Nair, A., & Parkar, S. (2019, April). Visual image caption generator using deep learning. In 2nd international conference on advances in Science & Technology (ICAST).

[11] Kesavan, V., Muley, V., & Kolhekar, M. (2019, October). Deep learning based automatic image caption generation. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

[12] Amritkar, C., & Jabade, V. (2018, August). Image caption generation using deep learning technique. In *2018 fourth international conference on computing communication control and automation (ICCUBEA)* (pp. 1-4). IEEE.

[13] Ding, S., Qu, S., Xi, Y., Sangaiah, A. K., & Wan, S. (2019). Image caption generation with high-level image features. *Pattern Recognition Letters*, *123*, 89-95.

[14] Oluborode, K., KADAMS, A., & Mohammed, U. (2024). An Intelligent Image Caption Generator Model Using Deep Learning. *International Journal of Development Mathematics (IJDM)*, *1*(3), 162-173.

[15] Padate, R., Jain, A., Kalla, M., & Sharma, A. (2023). Image caption generation using a dual attention mechanism. Engineering Applications of Artificial Intelligence, 123, 106112.