



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 9.935

Volume 15, Issue 10, October 2026

AI-Based Image and Video Captioning using Deep Neural Networks

Anusha B, Muskan, Pallavi H H, R Tamanna, Sahana V

Assistant Professor, Department of Computer Engineering, S J M I T Chitradurga, Karnataka, India

UG Students, Department of Computer Science Engineering, S J M I T Chitradurga, Karnataka, India

ABSTRACT: Visual data such as images and videos contain rich information, but understanding their content can be difficult for users who rely on text-based descriptions. This paper presents an AI-based approach for generating natural-language captions for images and video clips using deep neural networks. For image captioning, a convolutional neural network (CNN) extracts visual features from an input image, while a recurrent sequence model such as Long Short-Term Memory (LSTM) generates a sentence describing the detected scene. For video captioning, representative frames are sampled from the clip, visual features are extracted from those frames, and temporal information is combined to produce a concise description of the video. The proposed workflow includes input preprocessing, feature extraction, sequence generation, and caption post-processing. Public datasets such as MS COCO can be used for image-caption training and evaluation, while an appropriate video-caption dataset is required for video experiments. Performance should be assessed using established captioning metrics, including BLEU, METEOR, and CIDEr, alongside human review for relevance and readability. Quantitative results must be reported after implementation and testing on the selected datasets.

KEYWORDS: Image captioning, video captioning, deep learning, CNN, LSTM, computer vision, natural language generation, BLEU, METEOR, CIDEr.

I. INTRODUCTION

The rapid growth of digital media has resulted in large collections of images and videos across education, social media, healthcare, digital libraries, and surveillance. Although people can interpret visual scenes naturally, automatically converting visual content into meaningful language remains a challenging task. Image captioning aims to generate a sentence that describes the main objects, actions, attributes, and relationships in a still image. Video captioning extends this task by considering changes and events across a sequence of frames.

Traditional image-processing systems often rely on manually designed features and predefined rules. Such methods can struggle with complex scenes, varied lighting, occlusion, and diverse object relationships. Deep neural networks can learn useful visual representations from data and combine them with language-generation models. In a common encoder-decoder design, a CNN acts as the visual encoder and an LSTM acts as the language decoder. Attention-based and vision-language transformer models offer further ways to align relevant visual regions with generated words.

The objective of this project is to develop a practical pipeline that accepts an image or video and generates a concise natural-language caption. The system is organized into input handling, visual feature extraction, temporal feature aggregation for videos, caption generation, and output display. The remainder of this paper reviews related work, describes the proposed methodology, outlines evaluation procedures, and discusses limitations and future improvements.

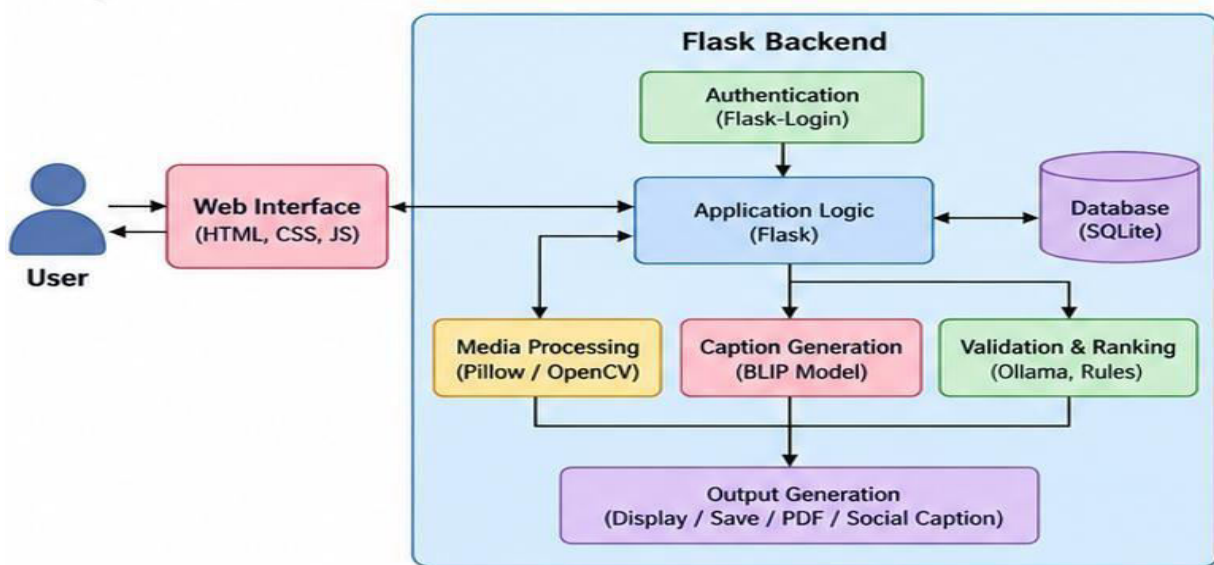
II. LITERATURE SURVEY

The Vinyals et al. introduced an encoder-decoder approach for image caption generation in which a convolutional network encodes an image and a recurrent neural network generates a sentence. Their work demonstrated that neural sequence models could produce captions directly from image features.

Karpathy and Fei-Fei explored the alignment between image regions and sentence fragments, helping establish methods for relating visual content to language. This line of research highlighted the importance of connecting words with the objects and regions they describe.

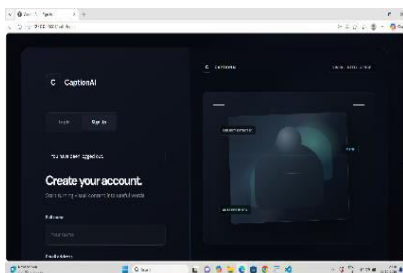
Xu et al. proposed an attention-based image-captioning model that allows the decoder to focus on different image regions while generating each word. Attention improves the model's ability to use localized visual information instead of relying only on a single global representation.

III. SYSTEM ARCHITECTURE

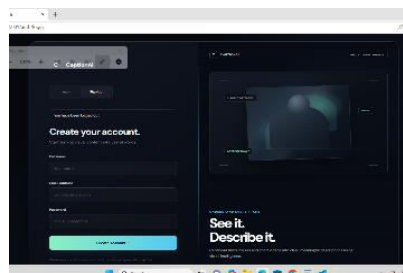


IV. EXPERIMENTAL RESULT

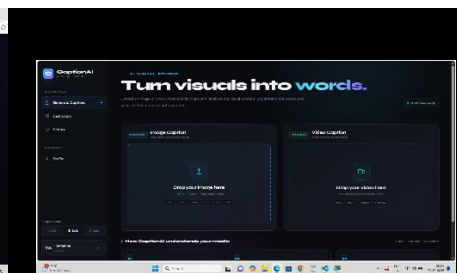
The experimental results demonstrate the performance of the proposed AI-based image and video captioning system using deep neural networks. The system processes input images and video frames to generate meaningful natural language descriptions of the visual content. The Convolutional Neural Network (CNN), such as ResNet, extracts important visual features from images, while the Long Short-Term Memory (LSTM) network generates descriptive captions based on the extracted features. For video captioning, multiple frames are analyzed to identify relevant objects, activities, and scenes. The generated captions are evaluated using metrics such as BLEU, METEOR, and CIDEr to assess their relevance and similarity to reference captions. The experimental results demonstrate the potential of deep learning techniques to generate accurate, meaningful, and contextually relevant captions for images and videos.



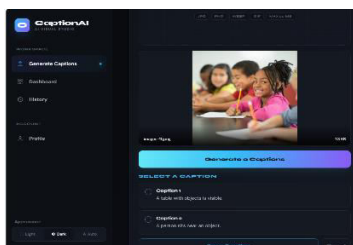
(a)



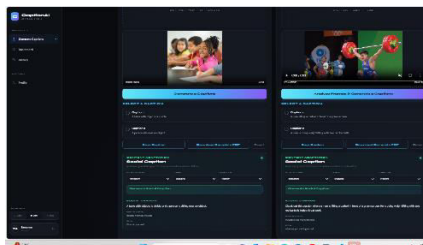
(b)



(c)



(d)



(e)

V. CONCLUSION

This paper describes a deep-learning pipeline for generating natural-language captions for images and videos. The proposed design combines CNN-based visual feature extraction with sequence-based language generation and extends the workflow to video through frame sampling and temporal feature aggregation. Its effectiveness must be established through experiments on suitable datasets and transparent reporting of BLEU, METEOR, CIDEr

REFERENCES

1. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
2. A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
3. K. Xu et al., "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," Proceedings of the International Conference on Machine Learning (ICML), 2015.
4. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997.
5. S. Venugopalan et al., "Sequence to Sequence—Video to Text," Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2015.
6. T.-Y. Lin et al., "Microsoft COCO: Common Objects in Context," European Conference on Computer Vision (ECCV), 2014.
7. C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," Text Summarization Branches Out: Proceedings of the ACL Workshop, 2004. (Background reference for automatic text evaluation; use caption-specific metrics for the final evaluation.)
8. R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-Based Image Description Evaluation," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details