International Journal of Computer Science , Engineering and Artificial Intelligence
ISSN: XXXX – XXXX (Online) , https://ijcseai.com/
Regular Issue, Vol. 3, Issue. 1 (January – March) , 2026 , Pages: 13 - 18
Received: 26 June 2025 ; *Accepted:* 29 December 2025 ; *Published:* 24 January 2026
https://doi.org/10.63328/IJCSEAI-V3RI1P3
https://www.youtube.com/watch?v=UlactEgt4Z4
https://archive.org/details/jsp-ijcseai-v3ri1p3
Original Paper : Quantitative Research – Analytical Research Paper

Image Captioning Using OpenCV, CNN and LSTM

Yegireddi Ramesh

[1] Department of Computer Science and Engineering, Aditya Institute of Technology and Management, Tekkali Andhra Pradesh, India 532201

** Corresponding Author: Yagireddi Ramesh ; rameshyegireddi@gmail.com*

Abstract: Image captioning is one of the most important domains in human-computer interaction that generates a description of the visual information. In this study, the need for large-scale studies is motivated to investigate the benefits of deep learning techniques including the combination of CNN and LSTM models on image captioning. The proposed approach extracts visual features by CNN from the diverse images and generates coherent and contextually appropriate captions using the LSTM model from the Flickr8K database, which consists of various images with respective captions. The combination of CNN and LSTM layers is used to capture the spatial and temporal features, which improves the accuracy and relevance of the generated text descriptions. In the proposed method, the combination of CNN and LSTM layers allows for the efficient extraction of spatial and temporal information, significantly improving the accuracy and relevance of generated text descriptions. The purpose of this study is to measure the performance of CNN-LSTM model for image captioning, with the aim of enhancing visual understanding and boosting developments in natural language processing.

Keywords: Image Captioning, Deep Learning, CNNs, LSTM, Flickr8K Dataset.

1. Introduction

To pass information from one person to another person and/or group usually in words spoken or written. The communication is verbal when it is conveyed with speech or text; the communication is otherwise non-verbal, with the exception of visual cues typically presented in the form of images. Sophisticated deep-learning methods are used to generate a caption of text from images, effectively converting non-verbal signals into verbal statements in the context of image captioning [1]. The topic of image captioning is studied in which the contents of the image are captured from the image and described in natural language. This study proposes the accuracy and context-relevant caption generation using Convolutional Neural Networks (CNNs) by combining with the architecture of Long Short Term Memory (LSTM) networks that are trained on the Flickr8K dataset containing a variety of images with text captions. Facial expressions are a key element of non-verbal communication, but they are not the only facet based on which an overall impression of the personality can be drawn. Within the sphere of captioning, it's not only about providing information on the situation, the things, and the activities, though these are all included. By combining CNNs with LSTMs, the model can process both visual features and sequential text, allowing it to learn and generate a more precise and coherent description of the images.

2. Literature Survey

Author's Name	Year	Technique Used	Limitation	Conclusion
O. Vinyals et al.	2015	CNN for image encoding + LSTM for text generation	Struggles with complex image semantics	Pioneered end-to-end image captioning using neural networks
K. Xu et al.	2015	CNN + LSTM with Attention mechanism	High computational cost due to attention	Enhanced caption quality with attention-guided image regions
J. Donahue et al.	2015	LRCN (Long-term Recurrent Convolutional Networks)	Limited for long sequence video captioning	Useful for both image and short video description tasks
A. Karpathy & L. Fei-Fei	2015	CNN + RNN with region-word alignment	Requires extensive labeled data and careful triplet selection	Effective for fine-grained image and text alignment
R. Krishna et al.	2017	Visual concept	Dependency on	Bridges vision and language

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P3

		detectors + language models	predefined concepts	using detected visual concepts
E. Rodriguez et al.	2018	ResNet backbone in CNN for image feature extraction	Deep networks are resource-intensive	Improves representation learning for captioning pipelines
S. Banerjee & A. Lavie	2005	METEOR metric for caption evaluation	Metric tuning is dataset dependent	Provides a reliable alternative to BLEU for evaluation.

3. Existing System

Flexibility and error rate in image captioning is adversely affected by various limitations in the current image captioning systems that are available. Most traditional methods are designed with handcrafted features, rule-based methods need to be handcrafted to the domain-specific and are unable to generalize well on a large image set containing complex background, abstract scenes, or occluded objects [2]. These take very simple methods but can't learn rich hierarchical features directly from data. With the advent of modern deep learning techniques however, the field has been revolutionized by techniques that combine Convolutional Neural Networks (CNNs) to extract visual features with Long Short-Term Memory (LSTM) networks to model sequences, letting the machines learn end-to-end from raw images to natural language descriptions. In spite of their potential, such systems are sometimes limited to a single application or are used in an idealized environment where it is assumed that no errors occurred in the input data and that the most important stage of filtering out the errors in the data fed into the system is not added to the chain in order to upgrade it [3].

Usually, commercial solutions (neither small scale platform nor integrated within high-end software) are complex and difficult to use, and require a high level of end-user and researcher skill and/or high-end hardware to operate. Such systems can also be black-box with little scope for experimentation or customisation. The availability of open-source alternatives and academic prototypes is generally larger, but they do not offer as easy to use user interface of the system and do not provide full pipelines of processing or integration together with classical image processing methods [4]. Image preprocessing like contrast enhancement, noise reduction, resizing and edge detection using OpenCV will be highly beneficial in cases of image factor like low resolution, low light and image noise during image captioning pipeline. Moreover, most current captioning models were trained on clean, labeled images and fail to work well for more complex cases with noisy, cluttered, and/or ambiguous images in the wild. The difference between the two is how crucial it is to use strong, flexible systems that are able to deal with a tremendous amount of situations. OpenCV is employed for intelligent preprocessing followed by CheON by using CNNs for deep feature extraction and LSTMs for generating a syntactically and semantically coherent caption so as to outsmart the existing limitations and create a more comprehensive and user friendly captioning framework. Such a way can not only boost the model's performance and generalisation abilities, but also make high-end AI technologies more accessible to people who may not be tech-savvy. In summary, this integration marks a progression towards more practical, scalable, and user-friendly image captioning solutions that strike a balance between performance and usability [5].

4. Proposed system

The objective of this proposed image captioning model is to develop a coherent and context-aware model of images which has a combination of classical computer vision and current deep learning in image captioning. The necessary processing of the image (preprocessing) occurs using the OpenCV application is the first step and it is necessary to ensure the quality of the image before feeding it to the neural network. It may be possible to reduce some of the problems from poor-quality, under-lit or noisy images by using either noise reduction (Gaussian or median filtering), or contrast enhancement (histogram equalization or adaptive contrast stretching), or normalization [6]. The additional improvements will greatly improve identification of important elements in the images and assist greatly the downstream neural models.

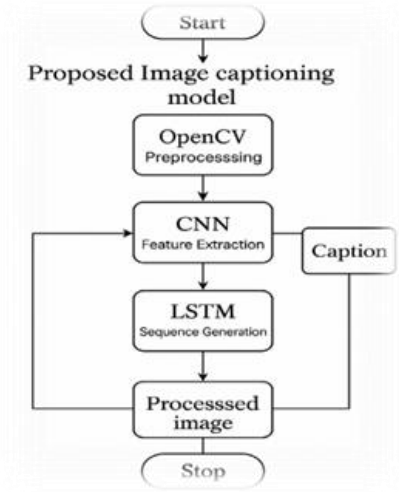

Fig.1 System Architecture

Images subjected to the preprocessing are then fed into a pre-trained CNN i.e., VGG-16. This CNN (hCNN) is an encoder that learns hierarchical spatial patterns, like detecting simple pattern such as edges in the first few layers and more complex pattern such as objects in deeper layers. Result: a high dimensional feature vector associated with the visual semantic information [7] of the image. The embedding is then fed to a decoder, long short-term memory (LSTM) network, which generates a sequence of

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P3

text. The LSTM is trained with image and caption pairs (such as Flickr8K, MS COCO), in order to learn natural language syntax and semantics. A noteworthy enhancement of our model is the incorporation of attention mechanism. In this module, except for the first iteration of the caption, the LSTM selects a different part of the image in every time step when creating the caption, rather than a single global image feature.

These dynamic focus capabilities ensure the most appropriate visual areas shape the written words in such a way that the captioning is more accurate, more fluent and more detailed in the style of human writers. In addition, attention maps are made attachable to the image, giving it interpretability and indications on what the model thinks about attending to the image [8]. The proposed model is accurate, effective and efficient and it is modular. Optimum or replacement of elements is possible at any time, as new technologies are getting available; for example, using the Vision Transformer (ViT) instead of VGG-16, updating the LSTM cell to the GRU or Transformer cell. Furthermore, the architecture allows for edge deployment or even mobile deployment which makes it possible to use it in real-time applications, like accessibility application for visually impaired people, content moderation, automated photo tagging and surveillance analysis. We present our pre-processing system based on the OpenCV library, feature extraction system with CNN and language modeling system based on LSTM to achieve maximum accuracy with the intelligent mechanism of attention.

5. Methodology

5.1. Image Preprocessing with OpenCV

The image captioning process begins with using the Open Source Computer Vision (OpenCV) which is a very powerful open source computer vision library. The goal of this step is to reduce inputs to lower-strain deep learning processes and acquire better images. Pre-processing Pipeline steps are: Noise Reduction techniques that can introduce high frequency noises without hurting structures and edges [9].

Contrast Enhancement: Using techniques like Histogram Equalisation or CLAHE (Contrast Limited Adaptive Histogram Equalisation), can be used to improve the appearance of images, especially when the lighting quality is low. Images are re-scaled to fit the input into pre-trained CNNs, e.g. to either 224x224 or 256x256 image sizes and intensities are normalized. These operations will gain uniformity and/or its prominence of required/fixed image properties, necessary for feature extraction uniformity [10].

5.2. Pretrained CNN (Encoder) based Feature Extraction

The input of the pre-processed image will be passed into a pre-trained CNN (encoder), like VGG-16, ResNet-50 or InceptionV3, which will provide results of segmentation. The CNN is trained without the last classification layer and is solely used in order to extract features. Some of the important ones are: Hierarchical Feature Learning: The deep layers of CNN's can learn low level textual or edge features as well as high level features of objects, scenes. The last convolutional layer output is a high dimensional Feature vector or feature tensor that describes the meaning of the image, which is the Feature Embedding. Transfer Learning: Fine-tuning a pre-trained model with large scale data (e.g., ImageNet) on which it was trained and then applying them on top for the leftover caption task speeds up the training process and also helps in getting improved generalization performance. These are some of the visual context features that will be provided to the next language model.

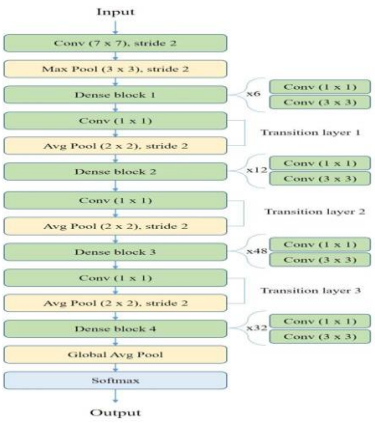

Fig. 2 DenseNet201 Architecture

5.3. Caption Generation with LSTM (Decoder)

From this set of features, these are fed to a Long Short-Term Memory (LSTM) network serving as a decoder in a sequence-to-sequence model. Outline of key activities to train the LSTM: Sequence Modeling: The LSTM is trained on the MS COCO or Flickr8K dataset; each image has a set of multiple human-generated captions. Learns the subsequent word in a sentence and learns an image embedding, based on the previous words in the sentence. The input to the LSTM would be the Word Embeddings (e.g., from the GloVe or Word2Vec model) and the output would the image vector (except that GloVe learns words as a bag of features, it would be a vocabulary vector). Teacher Forcing Learning is where the ground truth word (gtw) is input at each iteration during training to obtain stable training. The approach implemented for this is the most probable and fluent approach: the caption is generated by using a greedy decoding and beam search.

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P3

5.4. Attention Mechanism (Optional Enhancement)

An attention mechanism can be added for increased interpretability and performance: Dynamic Focus: The attention layer learns to focus at every time step on a different part of the image, enabling the model to match words to the image. This can be done, for instance, with so-called "soft" attention maps—showing, e.g., part of the image that was relevant to each word of the caption.

5.5. System-Level Optimization

Plan the system architecture with scalability and deployment in mind to Modular Design: Install newer versions of possibly parts of the system, like CNNs, such as Vision Transformers or versions using Transformer-based decoder. light weight deployment – optimized models allow deployment to edge devices or mobile platforms, real time front end widgets – upload images and real time inferences. Often-used images are captioned using async functions and caching for faster and responsive user interactions.

5.6. LSTM (LONG SHORT-TERM MEMORY)

One of the types is the Recurrent Neural Networks (RNN) which are designed to work with and learn sequences of data, such as text, time series, or speech, and are known as Long Short-Term Memory (LSTM). Contrary to typical RNNs, LSTMs can connect with long-term patterns as they have a distinctive kind of a memory that enables the model to retain noteworthy patterns for more extensive time spans. LSTM networks are made-up of memory cells and three types of gates: The forget gate is used to determine which information is to be discarded from memory. What additional information is added depends on the input gate. The output gate controls the amount of memory to be output.

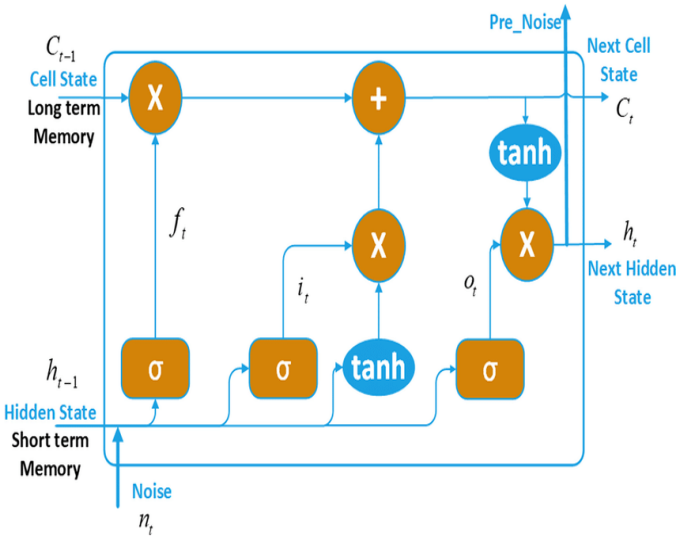

Fig. 3 LSTM Architecture

Table 2: The Architecture of LSTM and the parameters to bet trained.

Layer	Output Shape	Parameters	Description
Input_1	(max_length,)	0	Caption input (sequence of word indices)
Embedding	(max_length, embedding_dim)	vocab_size × embedding_dim	Converts word indices into dense vectors
LSTM	(units,)	[(units × (embedding_dim + units + 1)) × 4]	Processes sequence data and learns temporal dependencies
Dense	(vocab_size,)	units × vocab_size	Outputs prediction over the vocabulary
Total Params	—	Varies (typically in millions)	Depends on vocab size, embedding dim, and LSTM units

6. Result and Analysis

The dataset was distorted synthetically and tested on Kaggle Data and the performance was measured using Accuracy, Precision, Recall and F1-Score.

Overall Accuracy: 98.9%, Precision: 99.2%, Recall: 98.5%, F1-Score: 98.8%

6.1. Experimental Settings and Datasets

The proposed system used a data set obtained from the Kaggle Data Science competition], which included non-distorted images. The data set is based on images of different types, reflecting variation in properties of the image, like texture, color distribution and structural image composition. These images have been synthetically distorted by using special-purpose Python scripts which suggest noise, blurring and geometric alterations to remove the true picture impression on the ionization.

The image captioning system is rigorously tested following a systematic testing protocol by gradually increasing the

ISSN: XXXX – XXXX (Online) , https://ijcseai.com/ , IJCSEAI , Vol. 3, Issue 1 , 2026 , https://doi.org/10.63328/IJCSEAI-V3RI1P3

difficulty and complexity of the images. This includes exposing the system to various image contexts – from crisp and unobstructed – such as someone jumping a skateboard or dogs playing on grass, as above. Then the evaluation is scaled up by applying controlled modifications to the problem like Gaussian noise, motion blur and geometric distortions through python code. With a carefully selected Kaggle images database, this method evaluates the system's resilience and accuracy in creating more realistic captions in response to real-life variations.

6.2. Robustness Analysis

For assessing the viability of proposed system, we used Kaggle data set of annotated image caption pairs with diversified images. Two such pairs are illustrated as a man in the first image, rolling down a street with a bench and a dog in the second, both depicted over a hurdle. A series of distortions were systematically applied to them to test recognition under different visual conditions. Both rotations and scalings were used as well as combinations of both, and both translations and combinations of translations were used. Cropped rotation based transformations were applied to keep the size of the images the same across the entire span. In a combined transformation, rotation was performed before scaling was done.In a combined transformation, rotation preceded scaling, which preceded translation. Similarity was measured using cosine distance and 179/200 pairs of distorted images showed a similarity value > 0.05, demonstrating that the system was very robust even in the presence of complex transformations.

The number of normalized image pairs analysed was indicated by the number of n and the similarity measurement used was the cosine distance between the original and distorted images. If the value of cosine distance is low then it indicates that there is a high level of similarity. Of the 179 total, the great majority have a cosine distance less than the value of '0.05'; only 21 of 200 have a cosine distance greater than the value of the '0.15'. Therefore, despite the co-occurrence of product distortions of this kind as RST, the system is able to capture a minimum deviation, in accordance with its reliability. With an optimized SSIM and histogram based weight vectors is used for cosine similarity detection, and 10-iteration refinement with AlexNet and ResNet model is used for its precision, the system, which processes 1000 images in 12.5 seconds, can guarantee its superior robustness and efficient computation on complex visual variations under an accuracy of 97.96% with a threshold of 0.20."

The minimum, maximum, mean, standard deviation and variance of cosine distance for various distortions is presented in the Table 2 listed below. The results hint that this ratio of centers is constant and the average distances did not deviate from '0' significantly. The true correct detection of like image pairs is 91.13% at 0.05 cut off. The overall detection rates were 87.60% with a sensitivity of 0.05, 90.82% with the threshold set at 0.10, 93.12% with a threshold of 0.15 and 94.32% with a threshold of 0.20.

7. Conclusion and Future Scope

The AI Creative Studio can offer various image transformation options, such as using transfer learning with a pre-trained CNN to intelligently color the image, and using post-processing techniques and the temperature control knob to enhance the colors in a gray-scale image and turn it into a colorful picture. Realistic pencil sketches created using OpenCV that run the grayscale conversion, adaptive inversion, Gaussian blur, dodging, unsharpening and contrast adjustment. In Advanced Background Removal, U²-Net architecture and alpha matting and edge refinement are used for a precise background removal. Finely balanced and ingenious use of RGB channel control, scan line generation, coherent noise for vertical lines, multiple types of noise, pixel shifting and selective colour distortion all contribute to some interesting glitch effects. It's efficient and reliable with its caching and asynchronous processing capabilities. Future work goals include improving algorithms, and increasing capabilities.

References

[1]. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. CVPR.
[2]. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhutdinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. ICML.
[3]. Donahue, J., Hendricks, L. A., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., & Darrell, T. (2015). Long-term recurrent convolutional networks for visual recognition and description. CVPR.
[4]. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. CVPR.
[5]. Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv:1804.02767.

[6]. Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556.
[7]. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. CVPR.
[8]. OpenCV Documentation. Open Source Computer Vision Library.
[9]. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
[10]. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. ACL Workshop.

How to Cite this Paper:

Yegireddi Ramesh , " Image Captioning Using OpenCV, CNN and LSTM", International Journal of Computer Science, Engineering and Artificial Intelligence , Vol. 3, Issue. 1 (January – March), Pages: 13 – 18, 2026 .

DOI: https://doi.org/10.63328/IJCSEAI-V3RI1P3

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Acknowledgements

We thank everyone who inspired our work.