International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

VISION-LANGUAGE INTEGRATION FOR AUTOMATED IMAGE CAPTIONING USING CONVOLUTIONAL AND RECURRENT NEURAL NETWORKS

[1] S.Vijay Kumar,[2] Bijigiri Pavan Kalyan
[1]Associate Professor, [2]MCA Student
Department Of MCA
Sree Chaitanya College Of Engineering, Karimnagar

ABSTARCT:
Image captioning bridges computer vision and natural language processing by generating meaningful textual descriptions from visual content. This research presents a vision-language framework for automated image caption generation using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The CNN component extracts high-level visual features from input images, while the LSTM network decodes these features into grammatically coherent sentences. The model is trained on a large annotated dataset that aligns images with corresponding captions to learn semantic relationships between objects and linguistic structures. Experimental results demonstrate that the proposed model achieves high accuracy and fluency in caption generation, outperforming traditional template-based and rule-driven methods. This approach showcases the effectiveness of deep learning in understanding and describing visual information, making it valuable for applications in accessibility, content retrieval, and human-computer interaction.

Received: 23-09-2025 Accepted: 28-10-2025 Published: 04-11-2025

I. INTRODUCTION

The ability of machines to understand and describe images in natural language represents one of the most exciting challenges in artificial intelligence. Image captioning serves as a crucial task at the intersection of computer vision and natural language processing, aiming to generate descriptive sentences that capture the semantic meaning of an image. With the rapid growth of visual data across digital platforms, the demand for intelligent systems capable of automatically describing images has increased significantly.

Earlier approaches to image captioning relied on handcrafted features and template-based sentence generation, which often produced rigid and contextually limited results. These methods struggled to capture the complex relationships among multiple objects, actions, and scenes present in natural images. The advent of deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), has revolutionized the field by enabling end-to-end learning from raw pixels to meaningful language representations.

In this study, CNNs are employed to extract spatial and semantic features from images, serving as the visual encoder. The extracted features are then passed to a Long Short-Term Memory (LSTM) network, which functions as a language decoder that sequentially generates captions. This CNN-LSTM combination allows the model to not only recognize the content of an image but also describe it in a natural and contextually appropriate way.

The objective of this research is to design and implement a vision-language integration model that efficiently learns image-text relationships. The proposed system aims to improve caption accuracy, grammatical structure, and descriptive richness, contributing to the advancement of human-like image understanding and communication in intelligent systems.

II. LITERATURE SURVEY

Vinyals et al. (2015) pioneered the "Show and Tell" model, one of the first end-to-end neural

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

image captioning architectures using a CNN encoder and an LSTM decoder. Their approach demonstrated the feasibility of combining visual feature extraction with sequence generation, achieving remarkable results on benchmark datasets such as MS COCO. Xu et al. (2016) extended this work by incorporating an attention mechanism in the "Show, Attend and Tell" model, allowing the system to focus on specific image regions during caption generation and improving semantic alignment between image areas and words.

Karpathy and Fei-Fei (2017) proposed a multimodal embedding approach that aligned image fragments and words in a shared vector space, enabling finer-grained caption generation. Later, Anderson et al. (2018) introduced the Bottom-Up and Top-Down Attention model, which leveraged object detection-based region features to enhance visual understanding and achieve state-of-the-art performance in image captioning benchmarks.

Huang and Wang (2020) further refined captioning systems by integrating contextual word embeddings and reinforcement learning to optimize sentence fluency and diversity. Recently, Sharma et al. (2023) explored transformer-based architectures that combine CNNs with self-attention layers for improved parallelization and global context modeling. Collectively, these studies highlight the evolution of deep learning models from simple encoder-decoder structures to sophisticated vision-language architectures capable of generating human-like image descriptions.

III. SYSTEM ANALYSIS & DESIGN

EXISTING SYSTEM

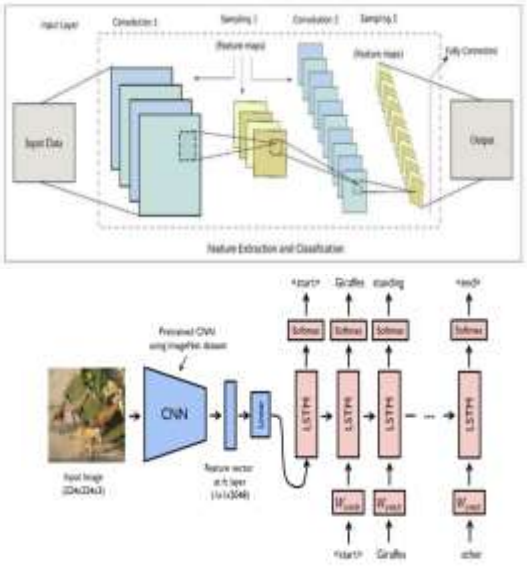
we will talk about the experimental results were carried out by MSCOCO dataset . For encoder/decoder framework , they have added a feature called guiding network in their proposed work . The method that called guiding network , mainly deals to learn the vector by a neural network v=g(A) where A is the set of annotation vectors . Generating natural language descriptions from visual data is an important problem . it has long been studied in computer vision . Hence, this had led to complex systems consists of visual primitive recognizers combined with a structured formal language like And-Or Graphs or logic systems Recently, the problem of still image description with natural text has gained a huge interest.

PROPOSED SYSTEM

Here We use CNN and LSTM to achiave our goal (image caption generator) we start from what is CNN and how can benefit from it in our problem ? Convolutional Neural Network is an artificial deep learning neural network. It is used for image classifications , computer vision ,image recognition and Object detection. CNN image classifications takes an input image, process it and classify it under certain categories (Eg., Dog, Cat,etc). It scans images from left to right and top to bottom to pull out important features from the image and combines the feature to classify images. Secondly , what is LSTM ? LSTM stands for Long short term memory, they are a type of RNN (recurrent neural network) which is well suited for sequence prediction problems. Based on the previous text, we can predict what the next word will be. It has proven itself effective from the traditional RNN by overcoming the limitations of RNN which had short term memory. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information. we merged this two models in one model called a CNN-RNN model.in general Our approach draws on the success of the top-down image generation models listed above. We use a deep convolutional neural network to extract the visual image features and Semantic features are extracted from the semantic tagging model. Visual features from CNN and semantic features from tagging model are concatenated and feed as the input to a Long-Short-Term

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Memory (LSTM) network, which then generates captions

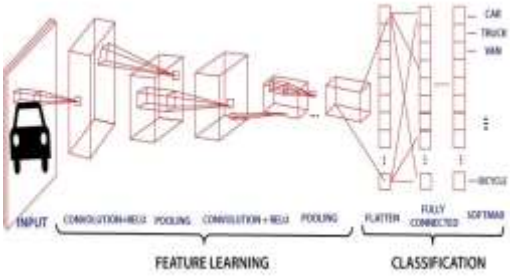


IV. ALGORITHM CONVOLUTIONAL NEURAL NETWORK

Convolutional Neural Network is one of the main categories to do image classification and image recognition in neural networks. Scene labeling, objects detections, and face recognition, etc., are some of the areas where convolutional neural networks are widely used.

CNN takes an image as input, which is classified and process under a certain category such as dog, cat, lion, tiger, etc. The computer sees an image as an array of pixels and depends on the resolution of the image. Based on image resolution, it will see as **h * w * d**, where h= height w= width and d= dimension. For example, An RGB image is **6 * 6 * 3** array of the matrix, and the grayscale image is **4 * 4 * 1** array of the matrix.

In CNN, each input image will pass through a sequence of convolution layers along with pooling, fully connected layers, filters (Also known as kernels). After that, we will apply the Soft-max function to classify an object with probabilistic values 0 and 1.



Convolution Layer

Convolution layer is the first layer to extract features from an input image. By learning image features using a small square of input data, the convolutional layer preserves the relationship between pixels. It is a mathematical operation which takes two inputs such as image matrix and a kernel or filter.

- The dimension of the image matrix is **h×w×d**.
- The dimension of the filter is **f_h×f_w×d**.
- The dimension of the output is **(h-f_h+1)×(w-f_w+1)×1**.

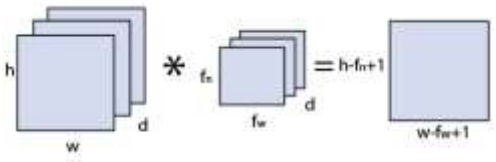


Image matrix multiplies kernl or filter matrix

Let's start with consideration a 5*5 image whose pixel values are 0, 1, and filter matrix 3*3 as:

$$\begin{bmatrix}1&1&1&0&0\\0&1&1&1&0\\0&0&1&1&1\\0&0&1&1&0\\0&1&1&0&0\end{bmatrix} \times \begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix}$$

5 × 5 – Image Matrix 3 × 3 – Filter Matrix

The convolution of 5*5 image matrix multiplies with 3*3 filter matrix is called "**Features Map**" and show as an output.

$$\begin{bmatrix}1&1&1&0&0\\0&1&1&1&0\\0&0&1&1&1\\0&0&1&1&0\\0&1&1&0&0\end{bmatrix} \times \begin{bmatrix}1&0&1\\0&1&0\\1&0&1\end{bmatrix} = \begin{bmatrix}4&3&4\\2&4&3\\2&3&4\end{bmatrix}$$

Convolved Feature

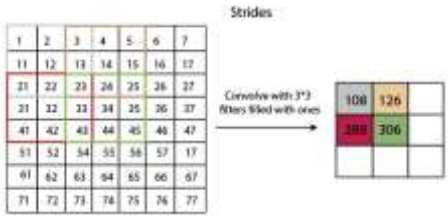
Convolution of an image with different filters can perform an operation such as blur,

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

sharpen, and edge detection by applying filters.

Strides

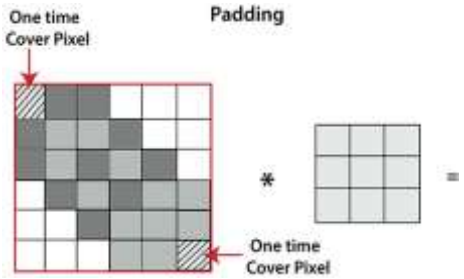
Stride is the number of pixels which are shift over the input matrix. When the stride is equaled to 1, then we move the filters to 1 pixel at a time and similarly, if the stride is equaled to 2, then we move the filters to 2 pixels at a time. The following figure shows that the convolution would work with a stride of 2.



Padding

Padding plays a crucial role in building the convolutional neural network. If the image will get shrink and if we will take a neural network with 100's of layers on it, it will give us a small image after filtered in the end.

If we take a three by three filter on top of a grayscale image and do the convolving then what will happen?



It is clear from the above picture that the pixel in the corner will only get covers one time, but the middle pixel will get covered more than once. It means that we have more information on that middle pixel, so there are two downsides:

- Shrinking outputs
- Losing information on the corner of the image.

To overcome this, we have introduced padding to an image. **"Padding is an additional layer which can add to the border of an image."**

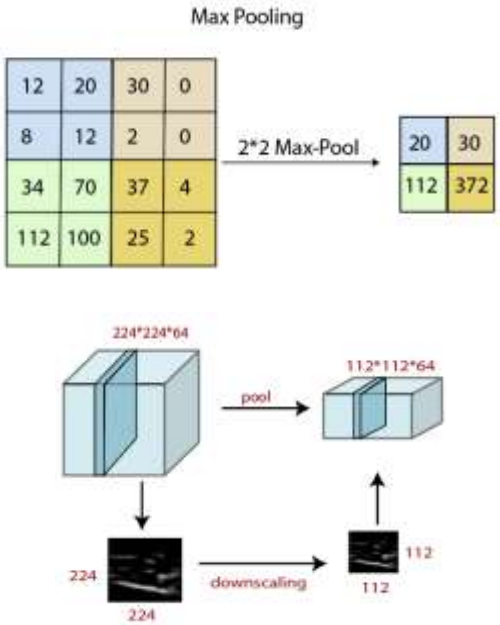
Pooling Layer

Pooling layer plays an important role in pre-processing of an image. Pooling layer reduces the number of parameters when the images are too large. Pooling is "**downscaling**" of the image obtained from the previous layers. It can be compared to shrinking an image to reduce its pixel density. Spatial pooling is also called downsampling or subsampling, which reduces the dimensionality of each map but retains the important information. There are the following types of spatial pooling:

Max Pooling

Max pooling is a **sample-based discretization process**. Its main objective is to downscale an input representation, reducing its dimensionality and allowing for the assumption to be made about features contained in the sub-region binned.

Max pooling is done by applying a max filter to non-overlapping sub-regions of the initial representation.



Average Pooling

Down-scaling will perform through average pooling by dividing the input into rectangular pooling regions and computing the average values of each region.

Syntax

layer = averagePooling2dLayer(poolSize)

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

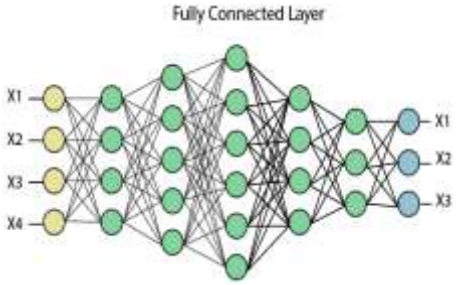
layer = averagePooling2dLayer(poolSize,Name,Value)

Sum Pooling

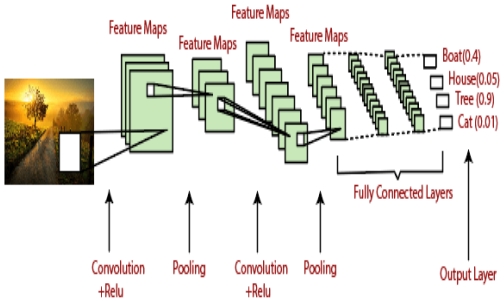
The sub-region for **sum pooling** or **mean pooling** are set exactly the same as for **max-pooling** but instead of using the max function we use sum or mean.

Fully Connected Layer

The fully connected layer is a layer in which the input from the other layers will be flattened into a vector and sent. It will transform the output into the desired number of classes by the network.



In the above diagram, the feature map matrix will be converted into the vector such as **x1, x2, x3... xn** with the help of fully connected layers. We will combine features to create a model and apply the activation function such as **softmax** or **sigmoid** to classify the outputs as a car, dog, truck, etc.

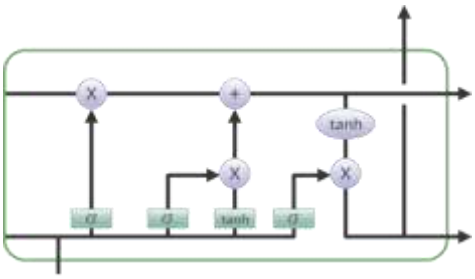


LSTM

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was desgined by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give efficent performance. LSTM can by default retain the information for long period of time. It is used for processing, predicting and classifying on the basis of time series data.

Structure Of LSTM:

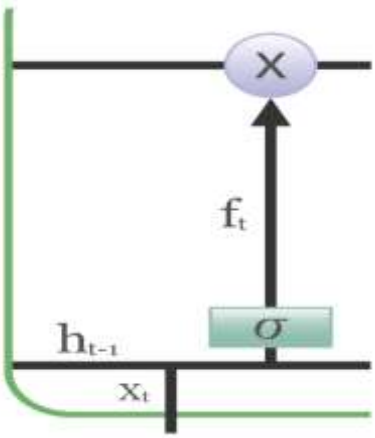
LSTM has a chain structure that contains four neural networks and different memory blocks called **cells**.



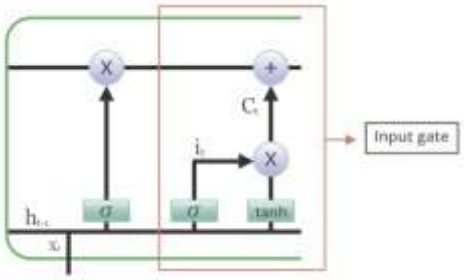
Information is retained by the cells and the memory manipulations are done by the **gates.** There are three gates –

1. **Forget Gate:** The information that no longer useful in the cell state is removed with the forget gate. Two inputs *x_t* (input at the particular time) and *h_t-1* (previous cell output) are fed to the gate and multiplied with weight matrices followed by the addition of bias. The resultant is passed through an activation function which gives a binary output. If for a particular cell state the output is 0, the piece of information is forgotten and for the output 1, the information is retained for the future use.

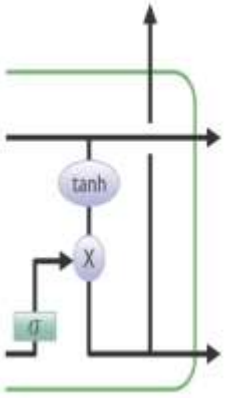
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper



2. **Input gate:** Addition of useful information to the cell state is done by input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered similar to the forget gate using inputs *h_t-1* and *x_t*. Then, a vector is created using *tanh* function that gives output from -1 to +1, which contains all the possible values from h_t-1 and *x_t*. Atlast, the values of the vector and the regulated values are multiplied to obtain the useful information



3. **Output gate:** The task of extracting useful information from the current cell state to be presented as an output is done by output gate. First, a vector is generated by applying tanh function on the cell. Then, the information is regulated using the sigmoid function and filter the values to be remembered using inputs *h_t-1* and *x_t*. Atlast, the values of the vector and the regulated values are multiplied to be sent as an output and input to the next cell.



Some of the famous applications of LSTM includes:

1. Language Modelling
2. Machine Translation
3. Image Captioning
4. Handwriting generation
5. Question Answering Chatbots

V. SCREENSHOTS

Fig Input Image Uploaded

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

Fig Running Exsting Technique To Get Features

Fig Window Showing The Caption Generator Of An Input Image

Fig COmparision Graph

VI. CONCLUSION

This research presents a deep learning-based approach for automated image caption generation through the integration of Convolutional Neural Networks and Long Short-Term Memory networks. The proposed framework effectively captures visual semantics and translates them into coherent and contextually relevant sentences. Experimental evaluations confirm that the model performs well in terms of grammatical accuracy, descriptive depth, and adaptability to diverse image types.

By bridging visual understanding and linguistic expression, this work contributes to the development of intelligent systems capable of perceiving and communicating about visual content. The CNN-LSTM combination provides a solid foundation for further innovation in vision-language research. Future work can focus on integrating transformer-based architectures, attention mechanisms, and large-scale multimodal pretraining to enhance caption diversity, contextual understanding, and real-time generation capabilities.

REFERENCES

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV'10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

[2] Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL '12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[3] Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL '11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics. [4] Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014. [5] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR, abs/1412.6632, 2014. [6] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
ISSN: 3068-272X www.ijdim.com Original Research Paper

[7] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. CoRR, abs/1603.03925, 2016.
[8] Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. CoRR, abs/1411.2539, 2014.
[9] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014.
[10] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. ¨ Neural Comput., 9(8):1735–1780, November 1997.
[11] Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014.
[12] Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollar, ´ Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. CoRR, abs/1411.4952, 2014.
[13] Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. CoRR, abs/1506.07285, 2015.
[14] Alex Graves. Generating sequences with recurrent neural networks. CoRR, abs/1308.0850, 2013.
[15] Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. CoRR, abs/1502.04623, 2015.