iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** II **Month of publication:** February 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.58298

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue II Feb 2024- Available at www.ijraset.com*

# Image Captioning:A Comprehensive Review

Yash Vardhan Mishra[1], Yashika Yadav[2], Shubham Kumar[3]
[1, 2]*Computer Science and EngineeringSRMCEM, Lucknow, India*
[3]*Professor, CSE, SRMCEM, Lucknow, India*

***Abstract:*** ***The Image Caption Generator is an intriguing project that bridges the domains of computer vision and natural language processing, aiming to automatically generate descriptive text for images. In the last decade, considerable progress has been made, yet key challenges in optimising Convolutional Neural Networks (CNNs) for precise image feature extraction and refining Long Short-Term Memory (LSTM) networks for coherent text generation persist. Challenges include overfitting, limited context understanding, and difficulties in modelling long-range dependencies. Solutions have emerged in the form of attention mechanisms, which help focus on relevant image regions, and advanced LSTM architectures like the use of Gated Recurrent Units (GRUs) to improve sequential modelling.***
***These innovations have significantly enhanced the overall performance, leading to more accurate and contextually relevant image captions, ultimately advancing the synergy between computer vision and natural language processing in the realm of image captioning.***
***Keywords: Image captioning, deep learning techniques, concepts of image captioning, CNNs, RNNs, LSTM, VGG-16***

## I. INTRODUCTION

Every day we encounter many visuals from various sources such as the internet, news, information, photographs and advertisements. These resources contain images that the viewer must interpret for himself. Most images do not have captions, but people can understand them in many ways without detailed captions.
But if people want automatic image tagging, machines need to define some kind of image signature. Signage is important for many reasons.
The names of all images on the web make image searches and evaluations faster, more descriptive and more accurate. Since scientists began studying the recognition of objects in pictures, it has become clear that simply naming the recognized object does not stop it from feeling like a good description for people. Unless machines think, talk, and behave like humans, natural explanations will be difficult to unravel.
Image signatures have many applications in many fields, including biomedicine, business, web search, and the military. Social media like Instagram and Facebook can create captions from photos. Creating labels for images is an important task in computer vision and natural language processing.
The ability of machines to copy human definitions is very important in the field of artificial intelligence. However, this approach does not have the necessary variables to create a rich word description.
With the increase in the power of neural networks, this limitation has been overcome. Most modern models use neural networks to generate text labels, take images as input, and predict the next word code in the resulting sentence.
.Following the introduction of the study in this section, Section 2 describes the literature review, while Section 3 explains the methodology. Section 4 presents the module description and their work,Section 5 discusses the architectural results of the study, and this is followed by conclusions and future directions in Section 6.

## II. LITERATURE REVIEW

Captioning has been attracting a lot of attention lately, especially in natural language. There is an urgent need for content based on the description of images, although this may seem a bit of an exaggeration, recent advances in fields such as neural networks, computer vision and natural language processing have contributed to the explanation. of images (i.e. representation of their visual effects). It means to step on the ground.
We achieve this by using advanced techniques such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) and images and their descriptions suitable for human perception. We found that our relationship model produced results in a regression experiment by the Flicker dataset.

©IJRASET: All Rights are Reserved 

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue II Feb 2024- Available at www.ijraset.com*

There are many types of video subtitles, some of which are rarely used today, but it is useful to give an overview of them before getting started.

1) *DEEP Learning-Based Image Captioning Methods:* Deep image captioning is categorized by modality, caption type, learning, and architecture, including attention-based, semantic, and LSTM.
2) *EN-DC Architecture VS. Compositional Architecture:* Some methods use just a simple vanilla encoder and decoder to generate captions. However, other methods use multiple networks for it.
3) *EN-DC Architecture-Based Image Captioning:* The neural network-based image captioning methods work in just simple end-to-end manner. This technique is similar to neural translation machines based on the encoder-decoder principle.
4) *Compositional Architecture-Based Image Captioning:* Compositional architecture-based methods composed of several independent functional building blocks:

First, a CNN is used to extract the semantic concepts from the image. Then a language model is used to generate a set of candidate captions.

*a)* Image features are obtained using a CNN.
*b)* Visual concepts (e.g. attributes) are obtained from visual features.
*c)* Multiple captions are generated by a language model using the information of Step 1 and Step 2.
*d)* CNN are used as encoders.
*e)* CNN are used in combination with RNN to analyze visual features.
*f)* CNN are used to extract features from images.

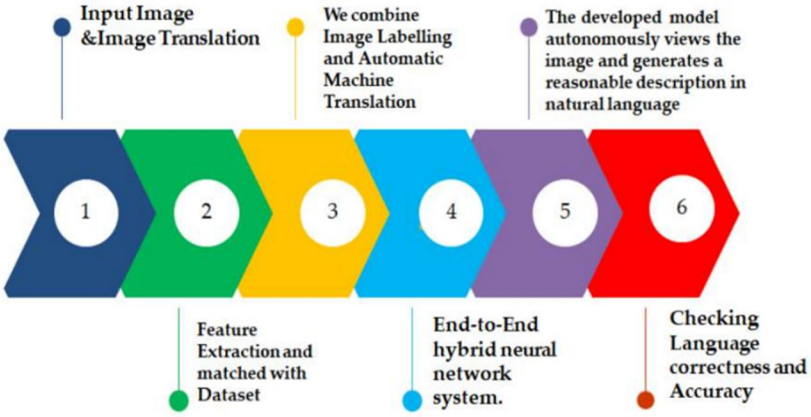
## III. PROPOSED METHODOLOGY

1) *Task:* The task is to create a system that will combine the image into a string and form a sentence
2) *Corpus:* We use the Flickr 8K dataset as the body. The file contains 8000 images, each consisting of 5 sentences. 5 tags for a picture help to understand all the events taking place. This file contains the Flickr_8k.trainImages.txt training dataset (6,000 images), the Flickr_8k.devImages.txt development dataset (1,000 images), and the Flickr_8k.testImages.txt testing dataset (1,000 images). The photos were selected from 6 different Flickr groups and do not include well-known people and places. However, they are manually selected to show a variety of scenes chosen among the selected captions and generate meaningful results.
3) *Data preprocessing:* The completed image and the relevant sentence in two places are cleaned and pre-processed. Image preprocessing is done by feeding the input data to the Xception application running on the Keras API running on top of TensorFlow. Xception was first trained on ImageNet. This helps us train images faster with the help of transformation learning. Annotations are cleaned up using the tokenizer class in Keras, which vectorizes the annotations and stores them in a separate dictionary. Each word in the table is then matched with a unique index value.
4) *Model:* Deep learning uses multi-layered, unstructured objects in a hierarchical structure to perform machine learning techniques. The model is based on a deep network where the data flow starts at the initial level and the model learns something simple and passes its output to the second layer of the network and combines the ideas with something more and moves to the third dimension. layer by layer. The process continues as each level in the network creates something more from the input it receives from the next level.

## IV. MODULE DESCRIPTION

1) Model Overview: The model is trained to obtain the probability p(SI); where S is a sequence of words generated by the model and each word is St. It is designed to use a dictionary created by teaching information. The input image is fed into a deep neural network (CNN), which facilitates the detection of objects in the image. The language is created as shown in Figure. Recurrent Neural Networks (RNN) take image encodings and use them to create image-related sentences. This model can be compared to the translation RNN model, where the goal is to maximize p(TS) and T is the translation of sentence S. However, in our model, the RNN encoder, which helps convert input sentences into long vectors, has been replaced by the CNN encoder.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



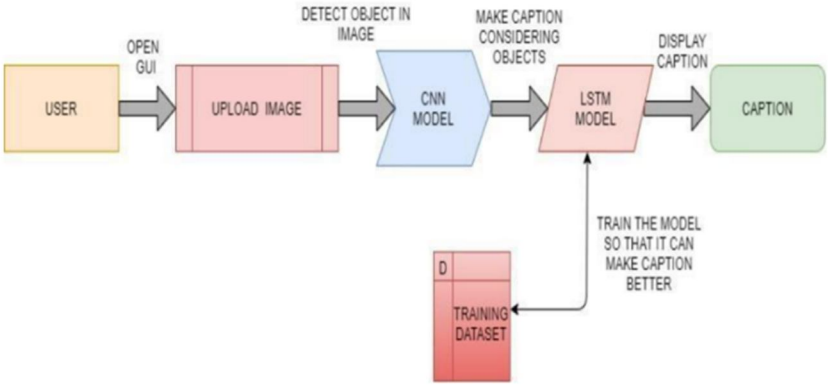

*2)* Recurrent Neural Network Feed-forward neural networks with internal memory are called recurrent neural networks. The result of the current input depends on the previous calculation; This makes RNNs circular as they do the same thing for every input. The output is created, copied, and sent back to the network loop.

*3)* Convolutional Neural Network Convolutional neural networks (CNN or ConvNet) are a type of deep neural network commonly used for visual analysis in deep learning. Translation-invariant and translation-invariant architectures based on the combination of weights are often called translation-invariant or location-independent artificial neural networks (SIANN). They can be used in many areas such as image and video recognition, recognition, image classification, image segmentation and medical image analysis. They can also be used in brain-computer interfaces, natural language processing, and financial time series. The multilayer perceptron was converted into a CNN.

*4)* Long-Term Memory Short-term memory (LSTM) networks are a type of recurrent neural network that can learn as expected in prediction problems. This is a good attitude to have for complex problems like machine translation and speech recognition.

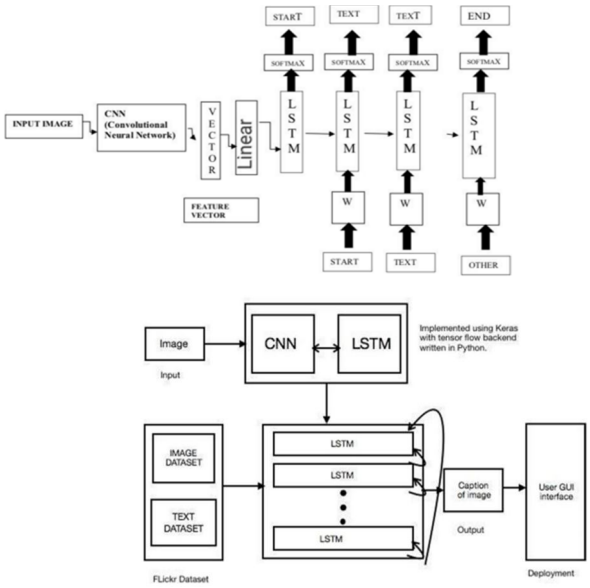
*5)* VGG16 It is considered one of the best visual architectures ever created. The most important part of VGG16 is the most important convolutional layer, which always uses the same padding and maxpool layer with a 3x3 filter in step 1 and a 2x2 filter in step 2. Convolutional and max-pooling layers are arranged in the same way throughout the architecture. Finally, there are two FCs (full coupling operation) followed by softmax for output.

The 16 in VGG16 means there are 16 weight layers. The network has over 138 million nodes, making it the largest network.




## V. ARCHITECTURE

The proposed model for the generator's signature diagram is shown in Figure 1 above. In this model, an input image is provided and a convolutional neural network is used to generate dense feature vectors as shown in the figure. This density vector is also known as embedding, which can be used as input for other algorithms and generate the necessary expressions for the output image. For the image signature generator, this signature becomes the representation of the image and is used as the initial state of the LSTM used to generate basic features for the image. Our system architecture is shown in Figure 2 below. This is what we want the system architecture to look like.




The image signatures creation process is responsible for creating captions for images provided during the course and can also create captions for new images. Our model uses image-based understanding, examines the image to identify objects present in the image, and produces text that describes the image so well that any machine can understand it. What is the picture trying to say?

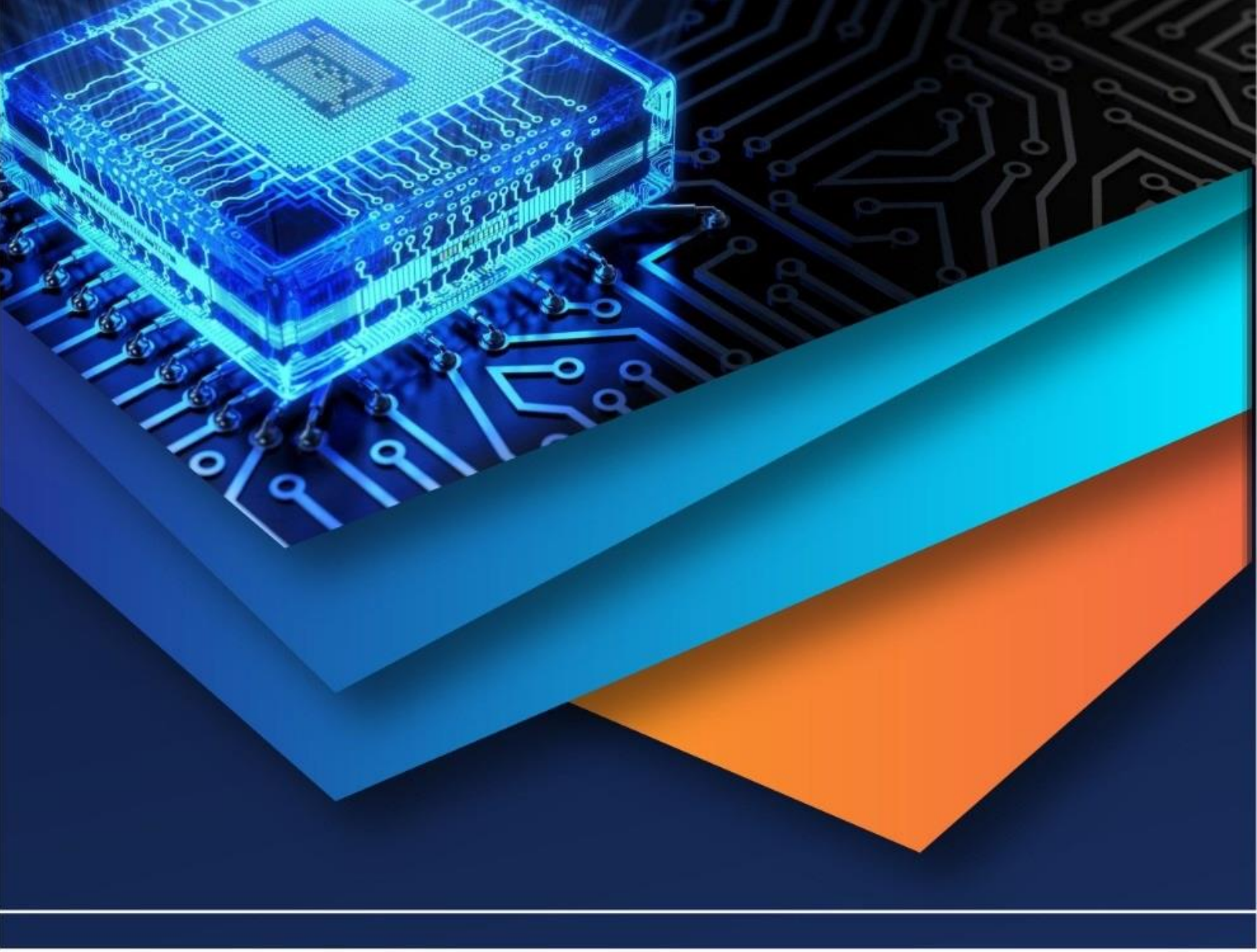
## VI. CONCLUSION

The image signature generator offers a wide range of applications across various industries by leveraging convolutional neural networks (CNN) and short-term neural networks (LSTM). A key area is accessibility, and these electronic devices play an important role in creating more digital content for people with visual impairments. They make the online experience more immersive by providing descriptions of images, allowing screen readers to share visual content accurately. In addition to ease of access, electronic signatures have also proven useful for indexing and searching. This tool is especially useful in e-commerce, stock photography, and photo libraries and helps search and organize large photo libraries. advanced detection. Social media platforms use visual signatures to enhance user experience and automatically create captions to provide context for images shared on platforms like Instagram and Twitter. Facilitates video content accessibility and search engine optimization (SEO) by including links to automatic video descriptions. In healthcare, graphic design aids in treatment and record keeping by aiding in the interpretation of medical images such as X-rays, MRIs and CT scans. Voice Recorder benefits include better accessibility, improved user experience, useful content, automation, and scalability. However, there are limitations in history. The main image at the time of captioning may make capture suitable for image distribution but not necessarily good for captioning. Optimization of the image encoder in the model can solve these problems and improve the capture of relevant information. Although the results are good, the generated names can sometimes be more revealing, indicating that the creativity and expression pattern can be improved. Future Work Subtitles have become an important issue in recent years due to the growth of social media and the internet. This report discusses various image retrieval research used in the past and introduces various techniques and applications in the research. Since feature extraction and similarity calculation in images are difficult in this field, there is a wide area for future research. Since these models do not depend on the content of the image, exact results are not possible. Therefore, a comprehensive study of the use of image content (e.g., image captions) for image retrieval will help solve this problem in the future. This work can be further developed in the future to improve the recognition of smaller groups by training with a larger dataset of signature images.

## REFERENCES

[1] S. ALBAWI and T. A. MOHAMMED, "Understanding of a Convolutional Neural Network," in ICET, Antalya, 2017.
[2] S. Hochreiter, "LONG SHORT-TERM MEMORY," Neural Computation, December 1997.
[3] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "A Neural Image Caption Generator," CVPR 2015 Open Access Repository, vol. XIV, 17 November 2014.
[4] D. S. Whitehead, L. Huang, H. and S.-F. Chang, "Entity-aware Image Caption Generation," in Empirical Methods in Natural Language Processing, Brussels, 2018.
[5] C. Elamri and T. Planque, "Automated Neural Image Caption Generator for Visually Impaired People," California, 2016.

[6] G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, "Neural Image Caption Generation with Weighted Training and Reference," Cognitive Computation, 08 August 2018
[7] J. Chen, W. Dong and M. Li, "Image Caption Generator Based On Deep Neural Networks," March 2018.
[8] S. Bai and S. An, "A Survey on Automatic Image Caption Generation," Neurocomputing, 13 April 2018.
[9] R. Staniute and D. Sesok, "A Systematic Literature Review on Image Captioning," Applied Sciences, vol. 9, no. 10, 16 March 2019.
[10] J. Hessel, N. Savva and M. J. Wilber, "Image Representations and New Domains in Neural Image Captioning," ACL Anthology, vol. Proceedings of the Fourth Workshop on Vision and Language, p. 29–39, 18 September 2015.
[11] M. Z. Hossain, F. SOHEL, M. F. SHIRATUDDIN and H. LAGA, "A Comprehensive Survey of Deep Learning for Image Captioning," ACM Journals, vol. 51, no. 6, 14 October 2018.
[12] A. Farhadi, M. Hejrati, M. A. Sadeghi and P. Young, "Every Picture Tells a Story: Generating Sentences from Images," in ACM Digital Library, 2010.
[13] S. Yan, F. Wu, J. Smith and W. Lu, "Image Captioning via a Hierarchical Attention Mechanism and Policy Gradient Optimization," LATEX CLASS FILES, vol. 14, 11 January 2019.
[14] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," CVPR 2015 Paper, December 2014.
[15] K. Simonyan and A. Zisserman, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," in ICLR, 2015.
[16] J. Donahue, L. A. Hendricks and M. Rohrbach, "Long-term Recurrent Convolutional Networks for Visual Recognition and Description," CVPR 2015, vol. 14, 31 May 2016.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)