iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VI **Month of publication:** June 2022

DOI: https://doi.org/10.22214/ijraset.2022.44502

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 10 Issue VI June 2022- Available at www.ijraset.com

Image Captioning Generator Using CNN and LSTM

M. Pranay Kumar[1], V. Snigdha[2], R. Nandini[3], Dr. B. Indira Reddy[4]
[1, 2, 3]*B. Tech (IV-IT),* [4]*Professor, Department of Information Technology, Sreenidhi Institute of Science and Technology, Hyderabad, Telangana, India*

Abstract: ***The project's goal is to come up with a caption for an image. Photograph captioning is the process of making a description for an image. It necessitates an understanding of the important things, their characteristics, and the relationships between them.an image's objects Deep learning techniques have progressed, and as a result, we can create models that can predict the future thanks to the availability of large datasets and computing power. Create captions for a picture This is what we've done in Python-based research in which we applied CNN's deep learning algorithm (Convolutional Neural Networks) and LSTM (Long Short-Term Memory) are two types of neural networks. Combining different types of RNNs (Recurrent Neural Networks) so that computer vision can be used A computer can recognize the context of a picture and present it in the appropriate context.***
Keywords: Image, Caption, Convolutional Neural Networks, Long short term memory, Recurrent Neural Network

I. INTRODUCTION

A complete human-like description makes a better impression. Natural language descriptions will remain a problem to be solved as long as machines do not think, talk, or behave like humans. Image captioning has various applications in various fields such as biomedicine, commerce, web searching and military etc. Social media like Instagram, Facebook etc. can generate captions automatically from images. Image captioning can give captions for both monochrome and color images of any pixel. Image caption generator is a task that involves computer vision and natural language processing concepts to recognize the context of an image and describe them in a natural language like English. In this Python based project, we will have implemented the caption generator using CNN (Convolutional Neural Networks) and LSTM a comprehensive human-like description makes a better first impression. Natural language descriptions will remain a difficult problem to address as long as machines do not think, talk, or behave like humans. Image captioning is used in a variety of sectors, including medical, commerce, web search, and the military.

Captions can be generated automatically from photographs on social networking sites like Instagram and Facebook. Image captioning may generate subtitles for any pixel in a monochrome or color image. The goal of creating an image caption generator entail using computer vision and natural language processing ideas to recognize the context of a picture and explain it in a natural language such as English.

The caption generator will be implemented using CNN Neural Networks in this Python-based project. We'll get the image features from Xception, which is a CNN model trained on the Flickr8k dataset, and then feed them into the features into the LSTM model that will be in charge of creating picture captions. Convolutional neural networks are a type of deep neural network that can analyze large amounts of data. It takes the form of a 2D matrix as input Images are simple to create. It is capable of handling the photographs that have been uploaded. Translated, rotated, scaled, and perspective shifts long short-term memory is abbreviated as LSTM. They're a form of RNN (recurrent neural network) that's good at predicting sequences issues. In light of the preceding sentence, we know what the following word is going to be. The LSTM can carry out relevant information throughout the processing of inputs, and it can discard non-related information using a forget gate.

II. LITERATURE SURVEY

In this section, we discuss the three main categories of existing image captioning methods: template-based image captioning, retrieval-based image captioning, and novel caption generation. Template-based techniques have fixed templates with blank slots to generate captions. In these systems, the different objects, actions and attributes are first identified and then the gaps in the templates are filled. For example, Farhadi et al. [1] use three different elements of a scene to fill the template slots for generating image captions. A Conditional Random Field (CRF) is leveraged by Kulkarni et al. [2] to detect the objects, attributes, and prepositions

©IJRASET: All Rights are Reserved

before filling in the blanks. Template-based approaches are able to generate grammatically correct captions, but since the templates are predefined, it cannot generate variable-length captions. In this section, we discuss the three main categories of existing image captioning methods: template-based image captioning, retrieval-based image captioning, and novel caption generation. Template-based techniques have fixed templates with blank slots to generate captions. In these systems, the different objects, actions and attributes are first identified and then the gaps in the templates are filled. For example, Farhadi et al. [1] use three different elements of a scene to fill the template slots for generating image captions. A Conditional Random Field (CRF) is leveraged by Kulkarni et al. [2] to detect the objects, attributes, and prepositions before filling in the blanks. Template-based approaches are able to generate grammatically correct captions, but since the templates are predefined, it cannot generate variable-length captions.

III. EXISTING SYSTEM

The most researched aspects of computer vision are image positioning and object detection. Users of social media platforms can now post photographs of any size or complexity and utilize Google to look for descriptions. Upgradeability, performance, flexibility, and scalability are all lacking. For input, high-quality images are required. In low-resolution photographs, features that are difficult to notice. Complex sceneries are tough to analyze. The goal of employing a proxy is to make the picture search process go faster. If the input image is intricate, it will take a long time to process, and you will be unable to post the gray scale image and also can't speak out the caption.

IV. PROPOSED SYSTEM

By providing appropriate, expressive, and fluid subtitles, Deep Neural Networks can tackle the problems that emerge in both versions. Accelerate the creation of subtitles. Users of social media will no longer have to waste hours searching for subtitles on Google with the system we offer. Our technology provides an easy-to-use platform for social network users to upload selected photographs. Uploading photographs does not require users to manually input captions. The proposed framework is capable of resolving the picture search issue. Color and black-and-white photos of any size can be uploaded and also can read the caption out in English. Tensor flows and algorithms can be used by neural networks to solve any problem and provide appropriate, expressive, and fluent subtitles. It is feasible to calculate automatic metrics efficiently. You won't have to waste time searching for captions because they'll be generated automatically.

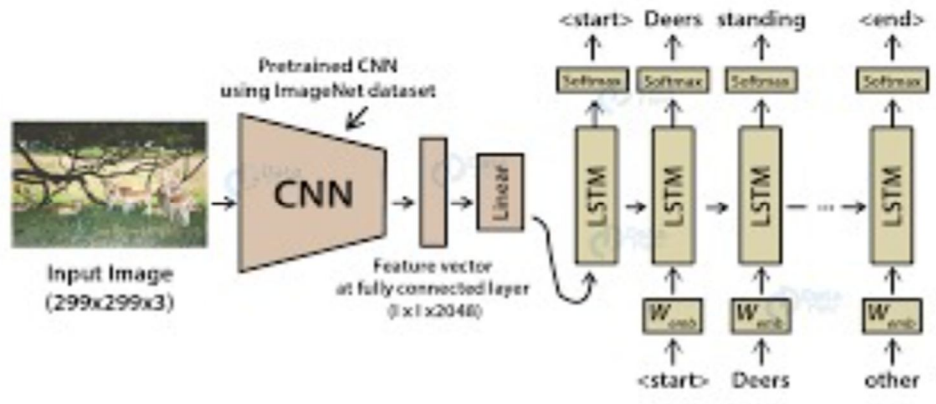
A. Task

The goal is to create a system that accepts an image in the form of a dimensional array, characterizes it, and provides syntactically and grammatically accurate statements as an output.

B. Corpus

As a corpus, I used the Flickr 8K dataset. The collection contains 8000 photos, each with five captions. A single image with five descriptions can help you grasp all conceivable circumstances. A training dataset Flickr 8k.trainImages.txt (6,000 photos), a development dataset Flickr 8k.devImages.txt (1,000 images), and a test dataset Flickr 8k.testImages.txt are all included in the dataset (1000 images).

V. WORKING MODEL



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue VI June 2022- Available at www.ijraset.com

VI. PROJECT ARCHITECTURE

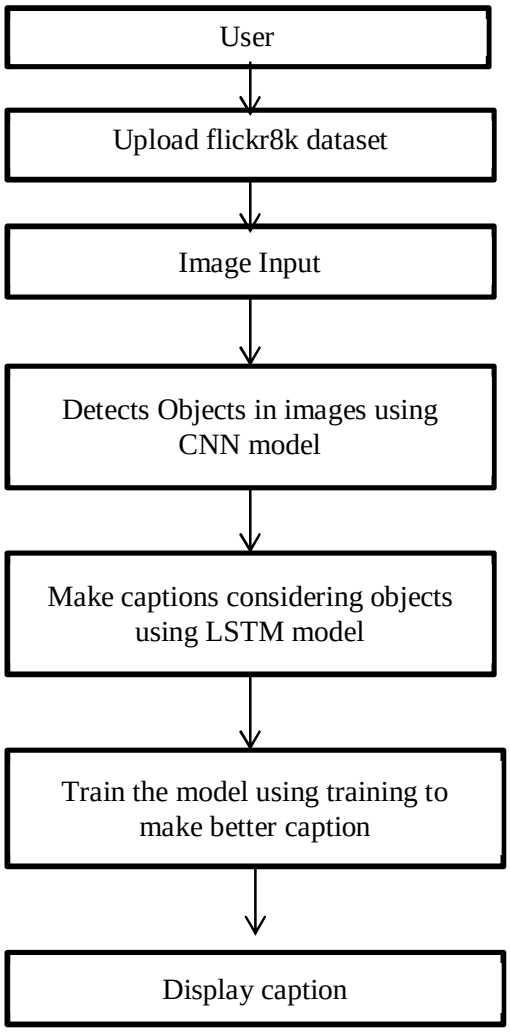


Figure: Flow Chart

VII. ALGORITHMS

A. Convolutional Neural Network

A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning system that can take in an image as input, assign importance (learnable weights and biases) to various aspects/objects in the image, and distinguish between them. When compared to other classification algorithms, the amount of preprocessing required by a ConvNet is significantly less. Convolutional neural networks are a type of deep neural network that can process data in the form of a 2D matrix. Images are easily represented as a 2D matrix, and CNN is an excellent tool for working with them. It scans photos from left to right and top to bottom to extract significant elements before combining them to classify them. In perspective, you can work with images that have been changed, rotated, scaled, and updated.

B. Long Short-Term Memory

LSTM stands for long short-term memory; it's a type of RNN (recurrent neural network) that's well suited to solving series prediction issues. We can guess what the next phrase will be based on the preceding paragraph. It has proven to be more powerful than traditional RNNs by overcoming the limitations of RNNs with short time period memory. LSTM may perform appropriate statistics while processing inputs, and it can discard non-applicable statistics via an overlook gate.

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

C. Flickr8k Dataset

The Flickr8k dataset is a publicly available image-to-set instruction benchmark. There are 8000 photos in this collection, each with five captions. These photos were gathered from several Flickr groups. Each caption includes a detailed description of the objects and events seen in the photograph. Because it depicts many events and settings and does not include photographs of renowned individuals or places, the dataset is broader. The training dataset has 6000 photos, the development dataset has 1000 images, and the test dataset has 1000 images. The following are the properties of the dataset that is appropriate for this project: • Having many labels for a single image makes the model more common and prevents it from being overfit. • The image annotation model can work in numerous picture categories thanks to different training image categories, which makes the model more resilient

VI. RESULT



```
[7]  !python3 '/content/drive/MyDrive/testing_caption_generator.py' -i '/content/drive/MyDrive/input/287152637-H.jpg'
2022-01-01 07:31:25.106892: E tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable d

start two boys play soccer on field end
```



```
'/content/drive/MyDrive/testing_caption_generator.py' -i '/content/drive/MyDrive/project images/istockphoto-1252455620-170667a-modified.jpg'
2022-04-10 16:11:54.044491: E tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable de

start dog is running on the grass end
```



```
[39] !python3 '/content/drive/MyDrive/testing_caption_generator.py' -i '/content/drive/MyDrive/Flicker8k_Dataset/1433142189_cda8652603.jpg'

2022-04-10 16:26:47.142811: E tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable de

start man is climbing up rock end
```

VIII. CONCLUSION

We looked at deep learning-Image Captioning approaches in this article. It demonstrated how to categories image annotation approaches, displayed a generic block diagram of the main groupings of, and highlighted the advantages and disadvantages of. We've broken down the benefits and drawbacks of each of the metrics and datasets. There's also a quick rundown of the experiment's findings. We briefly discussed the various research options that could be pursued in this area. Although deep learning-based image labeling systems have made significant progress in recent years, robust image labeling approaches that can create high-quality labels for practically every image have yet to be achieved. With the introduction of new deep learning network designs, automated captioning will remain a hot topic of research for some time. It makes use of the Flickr 8k dataset, which contains around 8000 photographs, as well as the captions, which are kept in a text file. Although deep learning-based image labeling systems have made significant progress in recent years, robust image labeling approaches that can create high-quality labels for practically every image have yet to be achieved. With the introduction of new deep learning network designs, automated captioning will remain a hot topic for a long time. The number of people using social media is growing every day, and the majority of them submit images, therefore the supply of captions will grow in the future. As a result, this project will be beneficial to them.

IX. ACKNOWLEDGMENT

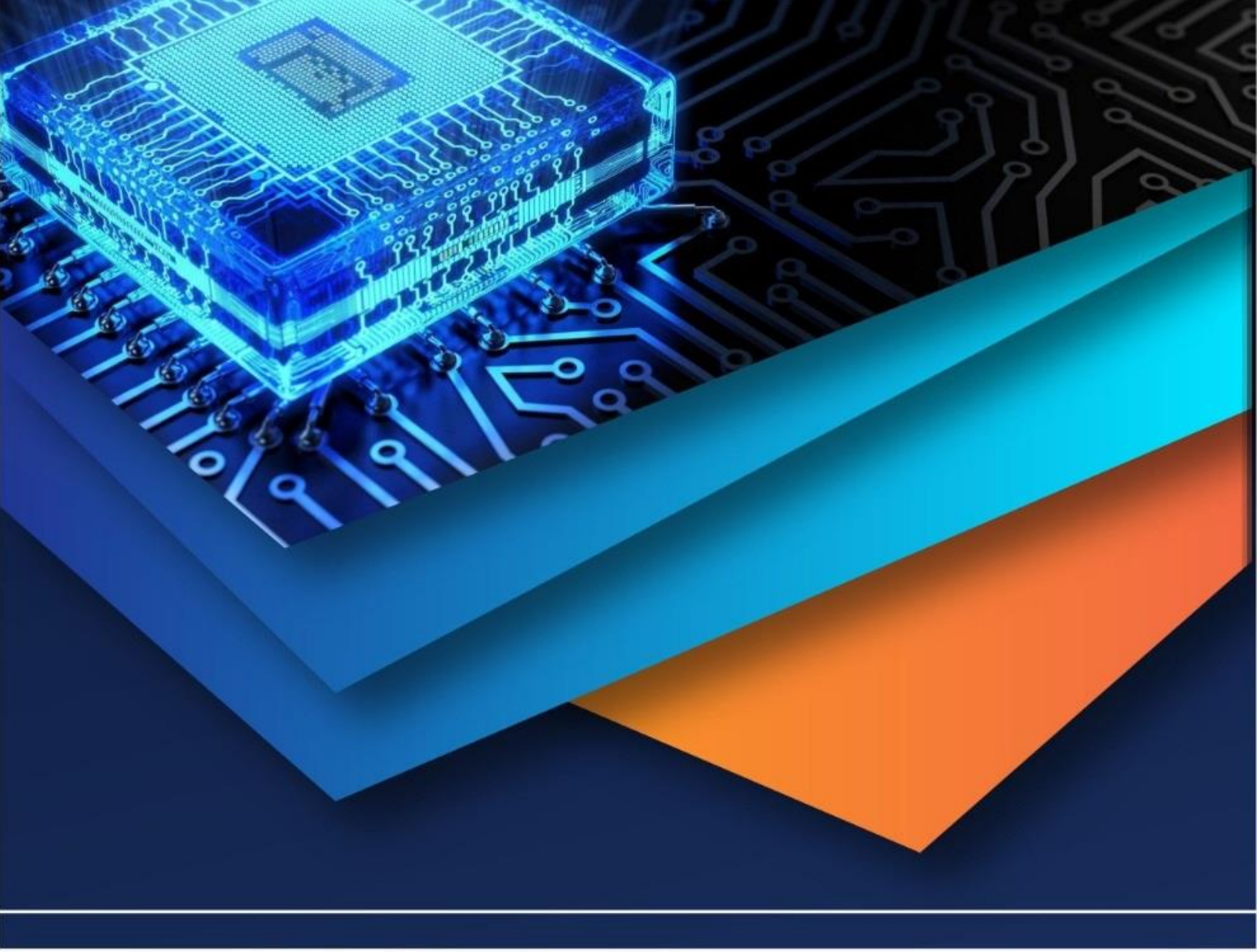
We would like to thanks to our guide Professor Dr. B. Indira Reddy and coordinator Professor Dr. M. Sreenivas for their continuous support and guidance. Due to their guidance, we can complete our project successfully. Also, we are extremely grateful to Dr. SUNIL BHUTADA, Head of the Department of Information Technology, Sreenidhi Institute of Science and Technology for his support and invaluable time.

REFERENCES

[1] William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation. arXiv preprint arXiv:1801.07736, 47, 2018.
[2] Girish Kulkarni, Visruth Premraj, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Baby talk: Understanding and generating image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35:2891–2903, June 2013.
[3] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings us- ing large weakly annotated photo collections. European Conference on Computer Vision. Springer, pages 529–545, 2014.
[4] Peter Young Micah Hodosh and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research, 47:853–899, 2013.

[5] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. Workshop on Neural Information Processing Systems (NIPS), 2014.
[6] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. International Conference on Machine Learning, 2048- 2057, 2015.
[7] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. IEEE International Conference on Computer Vision (ICCV), pages 4904–4912, 2017.
[8] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016.

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



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)