iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** IV **Month of publication:** April 2022

DOI: https://doi.org/10.22214/ijraset.2022.41589

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 IV Apr 2022- Available at www.ijraset.com

Gray Scale Image Captioning Using CNN and LSTM

V. Varshith Reddy[1], Y. Shiva Krishna[2], U. Varun Kumar Reddy[3], Shubhangi Mahule[4]
[1, 2, 3]*B. Tech (IV-CSE),* [4]*Assistant Professor, Department of Computer Science and Engineering, Ace Engineering College, Hyderabad, Telangana, India*

Abstract: ***The objective of the project is to generate caption of an image. The process of generating a description of an image is called image captioning. It requires recognizing the important objects, their attributes, and the relationships among the objects in an image. With the advancement in Deep learning techniques and availability of huge datasets and computer power, we can build models that can generate captions for an image. This is what we have implemented in this Python based project where we have used the deep learning techniques of CNN (Convolutional Neural Networks) and LSTM (Long short term memory) which is a type of RNN (Recurrent Neural Network) together so that using computer vision computer can recognize the context of an image and display it in natural language like English.*** *Gray Scale Image captioning can give captions for both monochrome and color images.*

Keywords: Image, Caption, Convolutional Neural Networks, Long Short Term Memory, Recurrent Neural Network

I. INTRODUCTION

Every day, we encounter a large number of images from various sources such as the internet, news articles, document diagrams and advertisements. These sources contain images that viewers would have to interpret themselves. Most images do not have a description, but the human can largely understand them without their detailed captions. However, machine needs to interpret some form of image captions if humans need automatic image captions from it. Image captioning is important for many reasons. Captions for every image on the internet can lead to faster and descriptively accurate images searches and indexing. Ever since researchers started working on object recognition in images, it became clear that only providing the names of the objects recognized does not make such a good impression as a full human-like description. As long as machines do not think, talk, and behave like humans, natural language descriptions will remain a challenge to be solved. 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. Gray Scale Image captioning can give captions for both monochrome and color images of any pixel. Gray scale 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 (Long short term memory). The image features will be extracted from Xception which is a CNN model trained on the Flickr8k dataset and then we feed the features into the LSTM model which will be responsible for generating the image captions. Convolutional neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easy. It can handle the images that have been translated, rotated, scaled and changes in perspective. 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. LSTM can carry out relevant information throughout the processing of inputs and with a forget gate, it discards non-relevant information.

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 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 retrieval-based methods, captions are retrieved from a pool of existing captions. Retrieval based methods initially find images that are visually similar images to the given image along with their captions from the training data set. These captions are called candidate captions. The captions for the given image are selected from this caption set [3], [4]. These types of systems produce general and grammatically correct captions. However, they cannot generate more descriptive and semantically correct captions.

Novel captions can be generated from visual and multimodal spaces. In these types of systems, the visual content of the image is first analyzed and then captions are generated from the visual content using a language model [5], [6], [7], [8]. These approaches can generate new, more semantically accurate captions for each image. Most novel caption generation techniques employ deep machine learning. Therefore, in this paper we focus primarily on deep learning based novel image caption generating methods.

Deep learning-based image captioning methods can also be classified based on learning techniques: Supervised learning, Reinforcement learning, and Unsupervised learning. We clump reinforcement learning and unsupervised learning into Other Deep Learning. Captions are most often generated for a whole scene in the image. However, captions can also be generated for different areas of an image such as in Dense captioning. Image captioning methods can either use simple Encoder- Decoder architecture or Compositional architecture.

III. EXISTING SYSTEM

The most popularly studied computer vision problem includes localization and object detection in images. Existing system allows the social media user to upload the image of their choice of any dimensions and having complexity and search for the caption in the Google. It lacks in updatability, performance, flexibility and scalability. Excellent image quality is required as input. Hard to detect features from low quality image. Difficult to analyze complex scene Usage of proxy is done for speeding up the image retrieval process. Time consuming when the input image is complex cannot upload Grey-scale images.

IV. PROPOSED SYSTEM

Deep neural network can solve the problems occurring in both the issues, by generating suitable, expressive and fluent captions. It speeds up the development process of image captioning. In the system proposed by us, social media user does not have to waste time on searching captions suitable for image on Google. Our system provides a user friendly platform for the social media user to upload the image of their

Choice. User doesn't have to type the caption manually for the uploaded image. Proposed framework can solve the image retrieval issues. Can upload both colour and black & white images of any dimensions. Neural networks can handle all the issues by generating suitable, expressive and highly fluent caption using tensor flow and algorithm. Efficient computation of automatic metrics is possible. As the captions are generated automatically there is no need to waste time on searching.

A. Task

The task is to build a system that will take an image input in the form of a dimensional array and generate an output consisting of a sentence that describes the image and is syntactically and grammatically correct.

B. Corpus

We have used the Flickr 8K dataset as the corpus. The dataset consists of 8000 images and for every image, there are 5 captions. The 5 captions for a single image helps in understanding all the various possible scenarios. The dataset has a predefined training dataset Flickr_8k.trainImages.txt (6,000images), development dataset Flickr_8k.devImages.txt (1,000 images), and test dataset Flickr_8k.testImages.txt (1000 images).

1) *Convolutional Neural Network:* A Convolutional Neural Network (ConvNet/CNN) is a Deep Learning algorithm which can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image and be able to differentiate one from the other. The pre-processing required in a ConvNet is much lower as compared to other classification algorithms. Convolutional Neural networks are specialized deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily represented as a 2D matrix and CNN is very useful in working with images. 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. It can handle the images that have been translated, rotated, scaled and changes in perspective.

2) *Long Short Term Memory:* 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.
3) *Flickr8k Dataset:* Flickr8k dataset is a public benchmark dataset for image to sentence description. This dataset consists of 8000 images with five captions for each image. These images are extracted from diverse groups in Flickr website. Each caption provides a clear description of entities and events present in the image. The dataset depicts a variety of events and scenarios and doesn't include images containing well-known people and places which makes the dataset more generic. The dataset has 6000 images in training dataset, 1000 images in development dataset and 1000 images in test dataset. Features of the dataset making it suitable for this project are: • Multiple captions mapped for a single image makes the model generic and avoids overfitting of the model. • Diverse category of training images can make the image captioning model to work for multiple categories of images and hence can make the model more robust.

V. PROJECT ARCHITECTURE

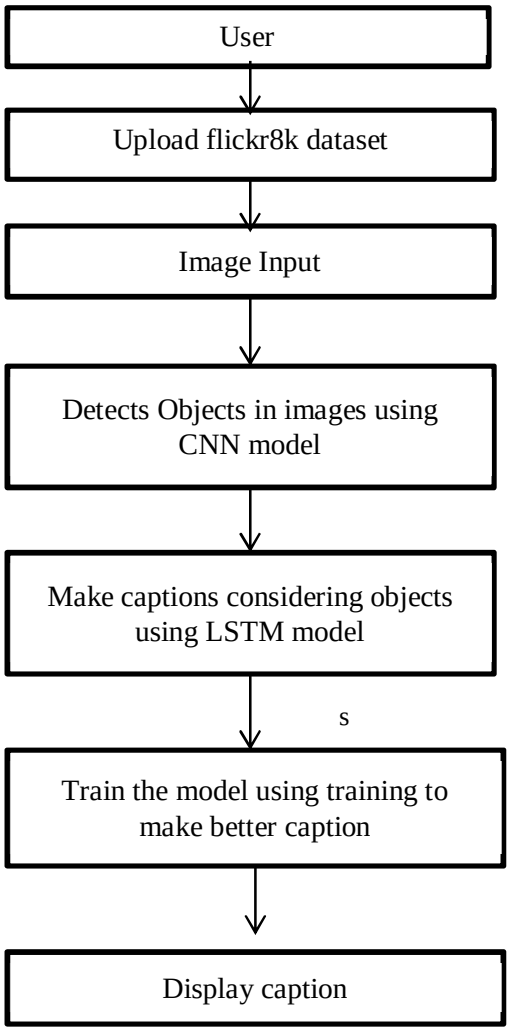


Figure: Flow Chart

VI. RESULT

```
[16] !python3 '/content/drive/MyDrive/testing_caption_generator.py' -i '/content/drive/MyDrive/project images/image.jpg'

2022-04-10 15:59:03.377659: E tensorflow/stream_executor/cuda/cuda_driver.cc:271] failed call to cuInit: CUDA_ERROR_NO_DEVICE: no CUDA-capable de

start two boys play soccer on field end
```

```
[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
```

VII. CONCLUSION

In this paper, we have reviewed deep learning-based image captioning methods. We have given a taxonomy of image captioning techniques, shown generic block diagram of the major groups and highlighted their pros and cons. We discussed different evaluation metrics and datasets with their strengths and weaknesses. A brief summary of experimental results is also given. We briefly outlined potential research directions in this area. Although deep learning-based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all images is yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for some time. We have used Flickr_8k dataset which includes nearly 8000 images, and the corresponding captions are also stored in the

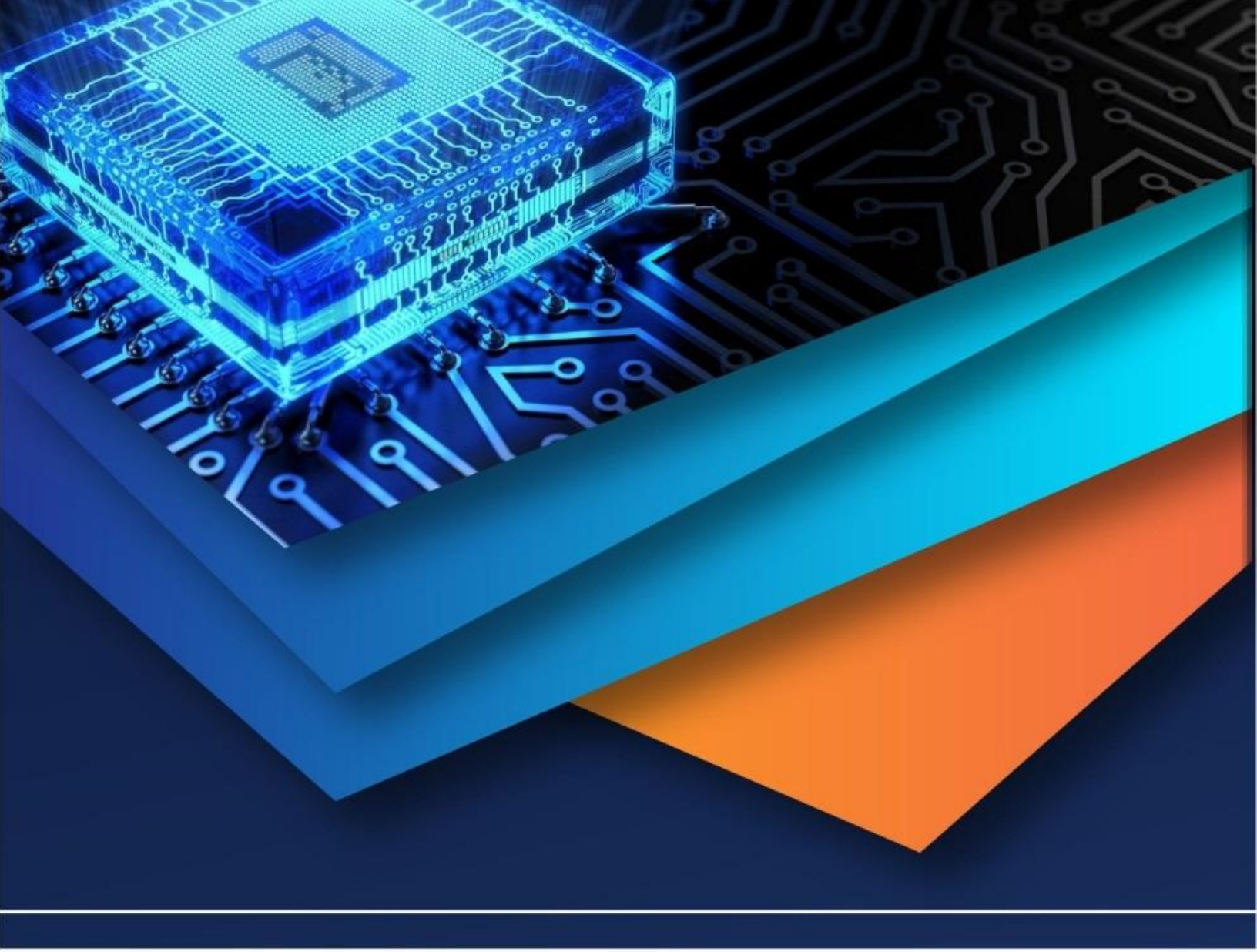
text file. Although deep learning -based image captioning methods have achieved a remarkable progress in recent years, a robust image captioning method that is able to generate high quality captions for nearly all
Images are yet to be achieved. With the advent of novel deep learning network architectures, automatic image captioning will remain an active research area for sometimes. The scope of image-captioning is very vast in the future as the users are increasing day by day on social media and most of them would post photos. So this project will help them to a greater extent.

VIII. ACKNOWLEDGMENT

We would like to thanks to our guide Assistant Prof. Mrs. Shubhangi Mahule and Associate Prof. Mrs. Soppari. Kavitha for their continuous support and guidance. Due to their guidance, we can complete our project successfully.
Also, we are extremely grateful to Dr. M. V. VIJAYA SARADHI, Head of the Department of Computer Science and Engineering, Ace Engineering College 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY