

# Image Synopsis Generating System Built using Deep Learning

Viksitha<sup>1</sup>, Ravi Kishan<sup>2</sup>, M T V Kumar<sup>3</sup>, Sai Klayan<sup>4</sup> and Babu<sup>5</sup>

<sup>1,3-5</sup>B.Tech ,V R Siddhartha Engineering College,CSE,Vijayawada,India

Email: viksithadirisana7981@gmail.com, {munhangiteja04, saikalyan521201, yarraguntababu117}@gmail.com

<sup>2</sup>Associate Professor, V R Siddhartha Engineering College,CSE,Vijayawada,India

Email: suraki@vrsiddhartha.ac.in

**Abstract**—When we saw an image, the human brain can easily understand the context of image, whereas can a computer know the context/content it representing? No, we make this possible in our project. Our work is to generate a sentence/description about the image automatically when it is given as a input. Our project model will take an image as an input and generate an english sentence as output, describing the contents of the image. The methods involved in the process of generating a caption are, one understands the content of the image done with deep learning algorithm like CNN-LSTM and to turn the understanding image into words. For this purpose, we use computer vision[6] and natural language processing. But, some people may not be able to understand the language displayed on the output screen and some may not be able to see. Hence, we convert this caption into the user's preferred language using a language translator (i.e. Google Translator) by importing gTTS and playSound libraries for text to speech conversion.

**Index Terms**— semantic, computer vision, natural language processing, CNN[1], LSTM[2], language translator, python Tkinter, Google translator, gTTS, playsound.

## I. INTRODUCTION

In recent times, image caption has bit by bit attracted the attention of numerous scientists in the field of artificial intelligence[10] and has become an intriguing and difficult assignment due to the rapid improvement of artificial intelligence. Image caption, consequently generating natural language depictions as per the content witnessed in an image, is a significant piece of scene understanding, which joins the knowledge of computer vision and natural language handling. The application of image caption exists extensive and important. The realization of human-computer relationships is one of the examples of an image caption[7] generator. The advancement of the image description device may help visually impaired individuals to “see” the world in the future. Currently, it has drawn growing attention and turn out to be one of the most vital subjects in computer vision. Prior the techniques for image description generation utilized in gathering image data through static object class libraries in the image and displayed utilizing factual language models. Consequently portraying the content of images utilizing natural language is a fundamental and challenging task. There is a lot of advancement in computing power along with the availability of huge datasets, building models that can generate captions for an image has become possible. Like humans, they can easily detect and express the environment they are in. Given a picture, it's natural for a person to explain a basic amount of details about the image with a rapid glance but the computer

needs help. Nowadays technology has been developing at a great rate and great development has been made in computer vision[6], tasks such as recognizing an object, action classification, image classification, attribute classification, and scene recognition is possible but it is a relatively new task to let a computer describe an image that is sent to it in the form of a human-like sentence.

Automation[9] is everywhere Manpower decreased due to involvement of machines as a result computational time cost decreased. Usage of these system later results into various applications such as surveillance watch man for security identifying vehicle numbers, we can use it in day-to-day life for education, practicing English or any other language, business, and travel. Image captioning is a task that requires a semantic understanding of images and the ability to generate description sentences with the proper and correct structure. Google Translate API provides a simple programmatic interface that is used in translating an arbitrary string into any supported language. This project requires to Google trans API it is used to detect and translate the language text between different languages. It is the most common application used by many people as it helps people who have the problem of understanding the language. Many people don't know all the languages as people mostly know only their respected language i.e. their mother tongue. Now the converted text is audible.

## II. LITERATURE SURVEY

Below sub-sections A, B, C, D is the literature survey for our paper.

### A. LSTM

Sepp Hochreiter[3], et.al explains about the deep neural network algorithm long short term Memory in "LONG SHORT-TERM MEMORY" Neural Computation, December 1997. In contrast with other algorithms, LSTM[3] leads to many more successful runs and learns much quicker. It is an artificial recurrent neural network (RNN) architecture utilized in the field of deep learning. Relative insensitivity to gap length is a benefit. Over RNNs, hidden Markov models, and other sequence learning strategies in various applications. The Advantage of it is solves even complex, artificial long-time lag tasks that has never been solved by a previous recurrent networks. The Disadvantages are that they require a lot of resources and time to get trained and become ready for real-world applications.

### B. Entity aware Image Caption Generation

D. S. Whitehead[2], et.al proposed a work in that they mentioned current image captioning approaches produce depictions they lack explicit data and here named elements are engaged with the images. In this work, the authors had proposed a new task that gives images as input which produces vivid image captions. Training a CNN-LSTM model that can generate a caption based on the image is a basic answer for this problem, so we have implemented this solution. The Advantage is that this model can incorporate real-world knowledge about entities across different modalities, generate text with better linguistic diversity. The Disadvantages are while training, the input of each time step comes from the real caption then generated captions each word is based on the previously generated word. If a word is not generated well, it may get far away from the ground truth. In our work, to overcome the disadvantages used the relative insensitivity of LSTM over RNNs, hidden Markov models, and other sequence learning strategies in various applications to solve the problems.

### C. Automated Neural Image Caption Generator for Visually Impaired People

C. Elamri[4], et.al stated that Automatically portraying the content of an image utilizing appropriately arranged English sentences is a challenging task, however, it is very much helpful for visually impaired individuals. At present smartphones are having many enhancing features, one of the features is clicking the photographs, which can help in taking encompassing pictures for visually impaired people groups. Here the image is given as input that generates captions loud enough so that it is beneficial mostly to the people who are visually impaired i.e they can hear and have an idea about what is happening around them. Here Christopher Elamri utilizes a CNN model to extract highlights of an image. These features are then fed into an RNN or an LSTM model to extract the portrayal of the image in linguistically correct English sentences depicting the environmental factors. The advantage is that The LSTM is apt to classify, process, and predict time series given time delays of unknown duration is the main advantage. The Disadvantages are that they require a lot of resources and time to get trained and become ready for real-world applications. In specialized terms, they need high memory data transmission because of linear layers present in every cell.

#### D. A Neural Image Caption Generator

O. Vinyals[5,10], et.al mentioned that the fundamental problem in artificial intelligence that connects computer vision and Natural language processing is automatically describing the content of images. Systematically analyze a deep neural networks-based image caption generation method. Here an image is given as input, and the method as output in the form of a sentence in English depicting the content of the image. The convolutional neural network model, recurrent neural network model, and sentence generation are three components of the system that are analyzed. The advantage is that this system analyzes the image and generates more trivial and relevant words for images. The disadvantage is that Due to its recurrent nature, the computation is slow. Training of RNN models can be difficult.

### III. BASIC CONCEPTS

The basic concepts of Python, Tkinter, gTTS and playsound libraries are used to implement this project. Tkinter is a standard GUI Python library. It is one of the fastest and easiest ways to build GUI applications. A python library that easily converts text into audio is known as gTTS (Google Text-to-Speech). playsound module is used to play audio files.

Datasets are the fuel for any machine learning algorithms because algorithms work on the data. Here Flickr Dataset is used which contains image dataset and text dataset. We collected the datasets from github and Thanks to Jason Brownlee [14,15] for providing a direct link to download the dataset (Size: 1GB). For the image caption generator, we have used the Flickr\_8K dataset , Flickr\_8k\_text .The Flickr\_8k\_text folder contains file Flickr8k.token which is the main file of our dataset that contains name of the image and their respective captions separated by newline("\n"). Flickr8k\_Dataset[15] folder contains 8091 images. Flickr\_8k\_text[14] Dataset folder contains the text files and captions of images.

Convolutional Neural networks are specialized deep neural networks[10] that can deal with the information that has an input shape like a 2D matrix used because helps in image classification, image recognition and feature extraction. CNN – LSTM used to create the new sequences in short and interprets the models in input sequence. One of the most popular deep neural networks is the convolutional neural network (CNN). There are multiple layers in CNN[1] such as a convolutional layer, nonlinearity layer, pooling layer, and fully-connected layer. A convolutional neural organization (CNN or ConvNet) is a part of deep learning mostly used in studying visual imagery, and also CNN is a class of deep neural networks. CNN's are regularized versions of multilayer perceptions. The Advantage of it has an excellent performance in machine learning problems and one of the most common algorithms. The Disadvantages are, in this, over fitting, exploding gradient, and class imbalance are the major challenges while training the model.

### IV. PROPOSED SYSTEM

The task of image caption generator includes computer vision and natural language handling concepts to perceive the context of an image and perceive them in a natural language like English. The image captioning is based on the semantics of images should be captured here and communicated in the desired form of natural languages. It has an incredible effect in reality, for instance by helping visually impaired individuals better understand the content of images on the web. Our image caption generator model can be achieved by merging CNN-RNN architectures. Feature extraction from images is done using CNN. We have used the pre-trained model Exception. The information received from CNN[1] is then used by LSTM[3,8] for generating a description of the image. However, sentences that are produced utilizing these methodologies are typically conventional depictions of the visual content, and foundation data is disregarded. Such generic descriptions do not fulfill emergent situations as they, basically replicate the data present in the images, and detailed portrayals with respect to events and entities present in the images are not given, which is basic to understanding emergent situations. We have utilized the Flickr\_8K dataset[15]t in the image caption generator project . There are also other big datasets like Flickr\_30K and MSCOCO dataset but it can take weeks for systems having only CPU support just to train the network, so we used a small Flickr8k dataset. Utilizing a huge dataset helps in developing a better model.

The project is developed as a web-based interface for users to get the depiction of the image and to make a classification system that can differentiate images as per their portrayal. It can likewise make the task of SEO simpler which is muddled as they need to keep up and investigate a huge amount of information. The generated caption will be taken as input by the language translator and by selecting the user's language the translator will convert it into the desired language. Computers can learn any language easily in a day whereas human beings

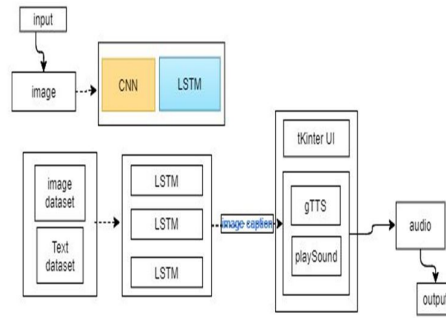


Fig-1 Basic System Architecture

take almost a year to learn a language. So this work's purpose is to make a way for users to translate text from one language to another using a mobile or web application. Instant language translator supports text translation in various languages.

Here in this model, input image is given & then a convolutional neural network is used to create a dense feature vector as shown in figure-2 System model. This dense vector, also called an embedding, this vector can be used as input into other algorithms, and it generates suitable caption for given image as output. For an image caption generator, this embedding becomes a representation of the image and used as the initial state of the LSTM for generating meaningful captions for the image. Now this generated caption is converted into user's language using google translator and the converted text is audible by using tkinter, gTTS, Playsound libraries in python.

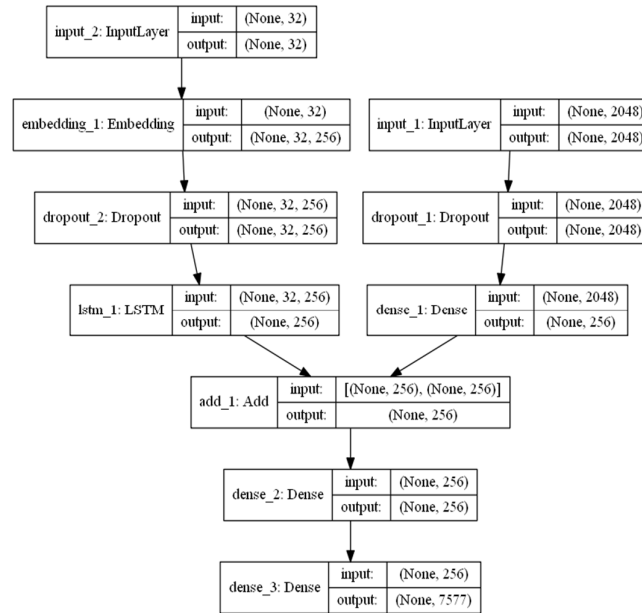


Fig-2 Image Synopsis System model

Convolutional Neural networks are very helpful to work with images because they are easily represented as a 2D matrix. Mostly the CNN[1] is utilized for image classifications and recognizing if an image is a bird, a plane, or Superman, etc. To pull out significant highlights from the image, it checks images from left to right and top to bottom and combines the feature to classify images. It can deal with the images that have been translated, rotated, scaled and alters in context. LSTM stands for Long short-term memory, which is a type of RNN (recurrent neural network) that is appropriate for sequence prediction issues. Based on the previous text, we can foresee what the following word will be. It has proven itself install effective from the traditional RNN by defeating the impediments of RNN which had short-term memory. LSTM[8] can carry out relevant information

throughout the processing of inputs and with a forget gate, it disposes of non-significant data. So, to make our image caption generator model, we will be consolidating these architectures. It is also known as a CNN-RNN model. The role of CNN is utilized for extracting features from the image. We will use the pre-trained model exception. LSTM will use the information from CNN to help generate a description of the image. The popular tkinter library for rendering graphics on a display window, googletrans library to translate text from one language to another. Google Translate API is free and it is helpful for experienced professionals in their business. Main advantage of Google Translate is that it is very fast. In fact, a human translator cannot compete with the neither speed nor as a result the quality of translations that Google Translate is able to perform. Google Translate API is able to produce a translation of text with same number of words in just seconds. In an average workday, an experienced Google translator can translate about 2000 words maximum depending on the difficulty of the text. This application will be useful to translate text and is best suited for single language usage. It is easy to translate the text. Googletrans library also has the feature to detect the language of the text. We used gTTs (google text to speech) library to convert text to voice and playsound library to play that converter voice from text.

## V. IMPLEMENTATION & RESULTS

Our project implementation requires a latest python version downloaded from the python website[16]. Pip installer to install all the mentioned modules in the proposed system section and the data set to train and test the data as software requirements and a operating system with basic configuration core i3 processor, 4gb RAM and 20gb Hard disk as hardware requirements. Open the program directory as shown in the below figure-4 to run the application. We need to enter the python program name and image path as shown below to run. As the image path for the figure-3 is "F://img1.jpg" the our model works perfect and gives the correct result as highlighted in the figure-4. It has taken nearly less computational time if you observe the below figure -4 the program implementation.



Fig-3 Boy is swimming in pool end

```
Microsoft Windows [Version 10.0.19041.985]
(c) Microsoft Corporation. All rights reserved.

C:\Users\hp\Desktop\major project\mini project>py testing_caption_generator.py -i "E://img1.jpg"
2021-06-25 10:07:22.184268: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'cudart64_110.dll'; this might be because the library is not placed in the expected location relative to the current executable
2021-06-25 10:07:22.184605: I tensorflow/stream_executor/cuda/cudart_stub.cc:29] Ignore above cudart dl error
2021-06-25 10:07:32.514330: W tensorflow/stream_executor/platform/default/dso_loader.cc:64] Could not load dynamic library 'cudart64_110.dll'; this might be because the library is not placed in the expected location relative to the current executable
2021-06-25 10:07:32.514442: W tensorflow/stream_executor/cuda/cuda_driver.cc:326] failed call to cuInit: UNKNOWN ERROR (303)
2021-06-25 10:07:32.519580: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:169] retrieving CUDA driver version is 11.0
2021-06-25 10:07:32.519827: I tensorflow/stream_executor/cuda/cuda_diagnostics.cc:176] hostname: 10.0.1.1
2021-06-25 10:07:32.522214: I tensorflow/core/platform/cpu_feature_guard.cc:142] This TensorFlow binary is optimized with oneAPI Deep Neural Network Library (oneDNN) to use the following CPU instructions in performance-critical operations: AVX AVX2
To enable them in other operations, rebuild TensorFlow with the appropriate compiler flags.
2021-06-25 10:07:34.726277: I tensorflow/compiler/mlir/mlir_graph_optimization_pass.cc:176] None of the optimizations listed above (including AVX and AVX2) were used.

start boy is swimming in pool end
```

Fig-4 Image Caption Display

As to help the people who are physically impaired and to help the illiterates the above display can be converted in to speech using translator on user preferred language from the list of the languages given around 180 languages our system works.

The output in figure-3 boy is swimming in pool end, when we select our required language then it translates that caption and makes it audible. Here we selected hindi as the language in figure-5 then it translates that caption into hindi language and makes it audible.



Fig-5 User Preferred Language

We have created a list of 180 languages for which google translator can work as Python list.

## VI. CONCLUSIONS

In this project, an image caption generator has been developed using a CNN-RNN model. Mainly the important aspect of our project is our model is dependent on the data so that it cannot predict the words that are out of its vocabulary. A dataset comprising of 8000 pictures is utilized here. In the case of the production-level models i.e. higher accuracy models, we need to train the model on larger than 100,000 image datasets so that better precision models can be created. User Interface is needed it has as importance as of working system it increases the scalability of users working. In future works use of flask/django to build web interface will be used.

## REFERENCES

- [1] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), pp. 1-6, 2017.
- [2] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, Shih-Fu Chang "Entity aware Image Caption Generation".
- [3] S. Hochreiter, Jurgen Schmidhuber, "Long short term memory", Neural Computation 9(8), pp. 1735-1780, 1997.
- [4] Christopher Elamri, Teun de planque "Automated Neural Image Caption Generator for Visually Impaired People".
- [5] O. Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "Neural Image Caption Generator".
- [6] M. Khademi and O. Schulte, "Image Caption Generation with Hierarchical Contextual Visual Spatial Attention," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 2024-20248, 2018.
- [7] P. Mathur, A. Gill, A. Yadav, A. Mishra and N. K. Bansode, "Camera2Caption: A real-time image caption generator," 2017 International Conference on Computational Intelligence in Data Science (ICCIDS), pp. 1-6, 2017.
- [8] J. Kim, J. Kim, H. L. Thi Thu and H. Kim, "Long Short Term Memory Recurrent Neural Network Classifier for Intrusion Detection," 2016 International Conference on Platform Technology and Service (PlatCon), pp. 1-5, 2016.
- [9] Y. Feng and M. Lapata, "Automatic Caption Generation for News Images," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 4, pp. 797-812, April 2013.
- [10] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and tell: A neural image caption generator," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3156-3164, 2015.
- [11] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj and R. K. Mishra, "Image Captioning: A Comprehensive Survey," 2020 International Conference on Power Electronics & IoT Applications in Renewable Energy and its Control (PARC), pp. 325-328, 2020.
- [12] S. Amirian, K. Rasheed, T. R. Taha and H. R. Arabnia, "Automatic Image and Video Caption Generation With Deep Learning: A Concise Review and Algorithmic Overlap," in IEEE Access, vol. 8, pp. 218386-218400, 2020.
- [13] V. Kesavan, V. Muley and M. Kolhekar, "Deep Learning based Automatic Image Caption Generation," 2019 Global Conference for Advancement in Technology (GCAT), pp. 1-6, 2019.
- [14] [https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k\\_text.zip](https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_text.zip).
- [15] [https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k\\_Dataset.zip](https://github.com/jbrownlee/Datasets/releases/download/Flickr8k/Flickr8k_Dataset.zip).