# Social Media Caption Generator Using Deep Learning Approach

**Nishad K. Shirsat[1], Prasad S. Patil[2], Aditya P. Chattikal[3], Prof Shweta Patil[4]**
[1, 2, 3, 4] Dept of Information Technology
[1, 2, 3, 4] PCE, Navi Mumbai, India - 410206

***Abstract-*** *When we want to caption an image or just describe it, it's very easy for humans to do that. But when we take a bulk of images it becomes very difficult and hectic to caption them, it also becomes very time consuming. So, to overcome this there are many caption generators used by many companies where a bulk of images can be captioned very easily. This process of captioning images is known as image captioning. Image captioning refers to the deep learning application of generating a textual description of an image using natural language processing (NLP) and computer vision (CV). This is how image captioning works. But we wanted to take this challenge one step further by generating captions automatically for social media pictures based on the image's features, looks, location, no. of people, etc. Social media captions tend to be more advanced than a simple descriptor and consists of puns, inside jokes, lyrics, references, sentiment, and some sarcasm. In some cases, the caption may not be relevant to the presented image at all. This System will help users to generate captions for their social media handles based on the photo they want to upload. The system helps people to find an appropriate caption for their image instantly and will eventually save time for users to manually find an appropriate caption for the image.*

***Keywords***- Image Captioning, Natural Language processing, Computer Vision.

## I. INTRODUCTION

We live in an era where everyone and mostly the young generation has a profile on some or other kinds of social media sites. We know how people like to post photos with different kinds of captions on their social media handles. People nowadays have a habit to post every detail of their life like going on vacation, going to colleges, hanging around with friends, spending time with family and friends, etc. through posting photos with a variety of captions on social media. But many times, it becomes a little time- consuming to find an appropriate caption for the image that needs to be posted. So our team came up with an idea where we could train datasets and generate captions for the images which could be then used to upload on social media handles. Image captioning refers to the Deep Learning application of generating a textual description of an image using Natural Language Processing (NLP) and Computer Vision (CV). This task requires an algorithm to not only understand the content of the image but also to generate language that connects to its interpretation. We wanted to take this challenge one step further by generating captions specifically for social media handles. Our work aims to generate captions that follow a specific style with specific vocabulary and expressions. To achieve this, we would require some basics of Deep Learning concepts such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), gradient descent, backpropagation, overfitting, probability, Python syntax, and data structures, Keras library, TensorFlow, etc. We will have to combine a Convolutional Neural Network for image classification with a Recurrent Neural Network for sequence modeling to create a single neural network that generates social media captions for images.

## II. LITERATURE SURVEY

**A.** ***A Systematic Literature Review on Image Captioning Published:*** Raimonda Staniut¯e and Dmitrij Šešok, it provides a brief overview of improvements which has been done in image captioning over the last four years. The main focus of this paper is to explain the most common techniques and the biggest challenges which are faced in image captioning. This paper is a systematic literature review of the newest articles available for image captioning. The databases used for this project is MS COCO and Flickr30K. the algorithms used for this technique is CNN for encoding and LSTM for decoding. The evaluation metric for this project used are BLEU-1, BLEU-2, BLEU-3, BLEU4 and CIDEr.

***B. Unified Vision-Language Pre-Training for Image Captioning and VQA:*** This paper presents a model called as Unified Vision language Pre-Training which is also known as VLP model. This project proposes a unified encoder-decoder model for general vision language pre-training. The paper uses a fine tuned pre-trained VLP model on the target dataset using the seq2seq objective. Datasets used for these models are COCO caption set, Flikr30k, and VQA 2.0. The model has been pre-trained on a large amount of text pairs using the unsupervised leaning objectives of two tasks: bidirectional and

sequence-to- sequence(seq2seq) masked vision language prediction.

***C. IMAGE CAPTION GENERATOR USING DEEP LEARNING:*** This paper uses Computer Vision (CV) and Natural Language Processing (NLP) of artificial Intelligence. It is being mainly used to help visually impaired people. The primary focus of this model is to recognize the image and then convert it to audio using GTTS. It uses LSTM to generate description of an image. The model used for this project is VGG-16. It has been successfully used to train and generate captions for images. The model basically depends on data so it cannot predict words that are out of its vocabulary. The model used for this project is implemented in KERAS framework. The accuracy score of this project is determined using BLEU. It uses CNN RNN model for image captioning.

D. ***Image Captioning - A Deep Learning Approach:*** This paper involves dual techniques from Computer Vision and Natural language Processing (NLP) to understand the image and turn the image into words. It uses algorithms like CNN to generate vocabulary and LSTM to generate sentences. It uses BLEU metric which is used to evaluate the model. The phases involved in this project are: 1. Image extraction, 2. Sequence processor and 3. Decoder. The database used for this project is flikr8k. The implementation of this project is done in Python SciPy Environment. The model used for this project is KERAS 2.0 which is a deep learning model. Here, TensorFlow library is installed as backend for KERAS framework which can be used for training deep neural network. The accuracy for this project is determined using BLEU.

## III. PROPOSED WORK

Social Media caption generator using deep learning approach is a system which will be used to generate automatic captions which can be used to post on social media.

**System Architecture**

The system architecture is given in Figure 1. Each block is described in this Section. Our network was inspired by Jason Brownlee's *How to Develop a Deep Learning Photo Caption Generator from Scratch* article.

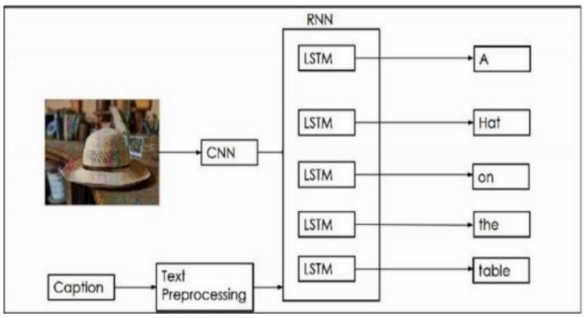


Fig. 1 Proposed system architecture

***A. Input Block Description:*** The first part is to upload an image in the web app that will be created. The user will have to upload an image file, which will be checked. If the uploaded file is not in the required file type (e.g. .jpg, .png, etc.) the file will be rejected. If the file is an image, then the file will be accepted and the image will be sent to the model for extracting the essential features which will then be used to generate the caption.

The model consists of 3 phases:

A. Image Feature Extraction

The features of the images from the Flickr 8K dataset is extracted using the VGG 16 model due to the performance of the model in object identification. The VGG is a convolutional neural network which consists of consists of 16 layer which has a pattern of 2 convolution layers followed by 1 dropout layers until the fully connected layer at the end. The dropout layers are present to reduce overfitting the training dataset, as this model configuration learns very fast. These are processed by a Dense layer to produce a 4096-vector element representation of the photo and passed on to the LSTM layer.

B. Sequence processor

The function of a sequence processor is for handling the text input by acting as a word embedding layer. The embedded layer consists of rules to extract the required features of the text and consists of a mask to ignore padded values. The network is then connected to a LSTM for the final phase of the image captioning.

C. Decoder

The final phase of the model combines the input from the Image extractor phase and the sequence processor phase using an additional operation then fed to a 256-neuron layer and then to a final output Dense layer that produces a SoftMax prediction of the next word in the caption over the entire

vocabulary which was formed from the text data that was processed in the sequence processor phase. The structure of the network to understand the flow of images and text is shown in figure 2.

4GB of RAM. It is also recommended to have a graphic card with a minimum of 2GB RAM and processing speed of up to 1650ti so that the dataset for building the model can be easily trained and it will also eventually save a lot of whenever we add more data to the dataset for training the model.

### 3.3 Dataset and Parameters

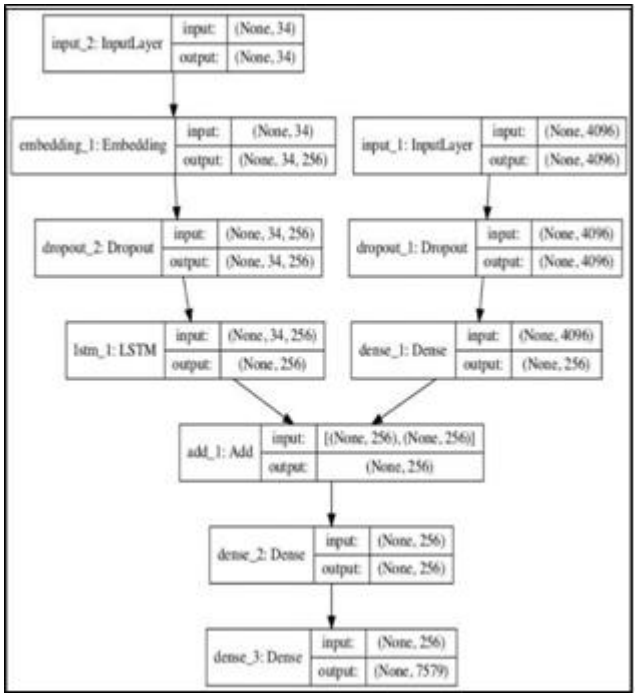

Figure 2. Image captioning model

## III. REQUIREMENT ANALYSIS

### Software

Firstly, the code for training the model for automatic captioning of images will be done in python IDE. We can use any IDE for writing python code like Spyder or Jupyter Notebook. We are going to train our model using deep learning algorithms like CNN, RNN LSTM. The CNN algorithm will extract the essential features from the image uploaded by the user, CNN is used to map image data to a particular output variable. Then we will apply an LSTM algorithm which will go through various phases like Image Feature Extraction, Sequence processor, and Decoder to extract the words from the images. The RNN LSTM algorithm will eventually combine the words and form a meaningful sentence from the image uploaded. Then the model which is trained using Kaggle dataset has to deploy on some web app or website. We can use streamlit to host the model on the web app where users can upload the image and get a specific caption for that image. We can also do the same thing using a website that can be developed using the python framework Flask.

### Hardware

There are no specific hardware requirements as such for this project. We will only be requiring a laptop or computer for running all the software used to build the required models for the project. It is recommended to have a processor of at least Intel i3 or Ryzen 3 and a minimum of
The Image captioning is a challenging dataset composed of 34,900 images from Kaggle. Then samples were uniformly distributed into three sets: the training, testing and validation.

## IV. CONCLUSION

Image captioning is a very exciting exercise and raises tough competition among researchers. There are more and more scientists who are deciding to explore this study field, so the amount of information is constantly increasing. It was noticed that the results are usually compared with quite old articles, although there are dozens of new ones, with even higher results and new ideas for improvements. With regards to our project which is social media caption generation where users can get captions for social media just by uploading a photo, there can be many things which might not work up to the expectations like sometimes caption might not be related to the uploaded image, sometimes if there might be some problems in data cleaning the captions might end up giving sentences which might not be in English or sometimes not generate any caption at all. Although the predicted results might lack diversity, they definitely would show that our model was able to learn and pick up the language and generate captions for social media users, As the training of the model with large datasets take a lot of time and computer processing. In future we can definitely use a supercomputer or more processing capacity computer wherein it will be possible to train large set of data in minutes and eventually it will result in better results and outputs for the project. Now we have tried to generate captions for users on social media in only one language i.e., English but in future we might work on generating captions for different languages as well. So, there is always ways to modify our model to improve the accuracy by using larger datasets, changing the model architecture and hyper parameter tuning (learning rate, batch size, etc.). However, as stated above, with our limited time and computing power, many of these modifications might be out

of reach for us. Despite all of our room for improvement, we can see our neural network is working fine, it did learn some great things and tried to generate some emoticon as well. Clearly, it still has a way to go, but we will hopefully inspire someone to continue and improve upon our work in the future.

## REFERENCES

[1] Staniut¯e, R., & Šešok *, D. (16 may 2019). A Systematic Literature Review on Image Captioning. 1-20. Department of Information Technologies, Vilnius Gediminas Technical University, Sauletekio al. 11, LT-10223 ˙ Vilnius, Lithuania; raimonda.staniute@stud.vgtu.lt Correspondence: dmitrij.sesok@vgtu.lt

[2] Zhou, L., 1, Palangi, H., 2, Zhang, L., 3, Hu, H., 4, Corso, J. J., 1, & Gao, J., 2. (4 dec 2019). Unified Vision-Language Pre-Training for Image Captioning and VQA. 1-10. 1 University of Michigan 2 Microsoft Research 3 Microsoft Cloud & AI 4 Microsoft AI & Research {luozhou, jjcorso}@umich.edu {hpalangi, leizhang, houhu, jfgao}@microsoft.com

[3] 1., Kousalya K., 2, Gokul S., 3, Karthikeyan R., 4, & Kaviyarasu D., 5. (2020). IMAGE CAPTION GENERATOR USING DEEP LEARNING. 1-6. 1Assistant Professor Department of CSE, Kongu Engineering College, Anna University, India 2Professor Department of CSE, Kongu Engineering College, Anna University, India 345UG student Department of CSE, Kongu Engineering College, Anna University, India

[4] Srinivasan, L. A., 1, Sreekanthan, D. A., 2, & L3, A. A., Undefined. (November 9, 2018). Image Captioning - A Deep Learning Approach. 1-4. 1,2 Student, Computer Science and Engineering, SRM Institute of Science and Technology 3Assistant Professor (O.G), Computer Science and Engineering, SRM Institute of Science and Technology