ASK IMAGE: A Chatbot which Answers Questions on Image Captions

Varun Vinod[1], Bhoyar Rohit Avinash[2], Jeetan Rajesh[3], Kale Gauresh Atmaram[4] and Prof. Dhiraj Amin[5]

[1-5]Department of Computer Engineering, Pillai College of Engineering, Navi Mumbai, India - 410206

---***---

Abstract - *Caption generation is a challenging artificial intelligence problem where a printed depiction should be produced for a given photo. It needs each strategies from pc vision to grasp the content of the image and a language model from the sphere of linguistic communication process to show the understanding of the image into words within the right order. A single end-to-end model is often characterized to anticipate a subtitle, given a photograph, instead of requiring refined data arrangement or a pipeline of expressly planned models. Conversational AI use cases are diverse. They include customer support, e-commerce, controlling IoT devices, enterprise, productivity and much more. In very simplistic terms, these use cases involve a user asking a specific question (intent) and the conversational experience (or the chatbot) responding to the question by making calls to a backend system like a CRM, Database or an API. It turns out that some of these use cases can be enriched by allowing a user to upload an image. In such cases, you would want the conversation experience to take an action based on what exactly is in that image. In this project we will develop a photo captioning deep learning model and incorporate COCO dataset, SQuAD dataset to provide rich and dynamic ML based responses to user provided image inputs.*

Key Words: **Natural Language Processing, Caption Generation, Convolutional Neural Networks, Recurrent Neural Networks, Question Processing.**

1. Introduction

Ask Image's fundamental target is to produce subtitles by preparing the info picture and coordinating a chatbot. Caption generation is an interesting artificial intelligence problem where a descriptive sentence is generated for a given image. It includes the double procedures from PC vision to comprehend the substance of the picture and a language model from the field of normal language preparing to transform the comprehension of the picture into words organized appropriately. Image captioning has various applications such as recommendations in editing applications, usage in virtual assistants, for image indexing, for visually impaired persons, for social media, and several other natural language processing applications. At the point when people read an article or a short entry from book, the most ideal path for checking a nature of far reaching perusing is attempting to make a rundown or responding to the inquiries with regards to the part that you read. Therefore in order to mimic this reading process most of the QA systems are aimed to extract important information from a provided article or a short passage to answer the given questions.

2. Literature Survey

2.1 Enriching Conversation Context in Retrieval-based Chatbots.

This technique is demonstrated by Amir Vakili and Azadeh Shakery from the University of Tehran. This project works on retrieval-based chatbots, like most sequence pair matching tasks, can be divided into Cross-encoders that perform word matching over the pair, and Bi-encoders that encode the pair separately. Development of a sequence matching architecture that utilizes the entire training set as a makeshift knowledge-base during inference is expanded upon. Retrieval-based systems, which select a response from candidates retrieved from chat logs according to how well they match the current conversation context as opposed to generative systems which synthesise new sentences based on the context are studied. Detailed experiments demonstrating that this architecture can be used to further improve Bi-encoders performance while still maintaining a relatively high inference speed are performed.

2.2 Survey on Automatic Image Caption Generation.

This survey is executed by Shuang Bai and Shan An for image caption generation. The survey explains in detail about connecting both research communities of computer vision and natural language processing. In this paper, a survey on advances in image captioning research based on the technique adopted and classification of image captioning approaches into different categories is presented. Representative methods in each category are summarized, and their strengths and limitations are talked about. The initial methods discusssed are mainly retrieval and template based. Neuural network based methods are also discussed, which give state of the art results. Neural network based methods are further divided into subcategories based on the specific framework they use. Each subcategory of neural network based methods are discussed in detail. After that, state of the art methods are compared on benchmark datasets. Following that, discussions on future research directions are presented.

2.3 An Intelligent Behaviour Shown by Chatbot System.

The Authors Vibhor Sharma, Monika Goyal , Drishti Malik discuss about how chatbots are software agents used to interact between a computer and a human in natural language, just as people use language for human communication, chatbots use natural language to communicate with human users. In this paper, analysis of some existing chatbot systems namely ELIZA and ALICE is observed. Arrival at a conclusion that it is easier to build bots using ALICE because of its simple pattern matching techniques that building one for ELIZA since it is based on rules is observed. Finally, discussion of the proposed system in which the implementation of ALICE chatbot system as a domain specific chatterbox which is a student information system that helps users in various queries related to students and universities is observed.

2.4 Chatbot Design-Reasoning about design options using i* and process architecture.

The Authors Zia Babar , Alexei Lapouchnian, Eric Yu discuss about how software systems are often designed without considering their social intentionality and the software process changes required to accommodate them. This paper considers chatbots as domain example for illustrating the complexities of designing such intentional and intelligent systems, and the resultant changes and reconfigurations in processes. A mechanism of associating process architecture models and actor models is presented. The modelling and analysis of two types of chatbots, retrieval based and generative are shown using both process architecture and actor models.

3. System Architecture

The system architecture is given in Figure 1. Each block is described in this Section.

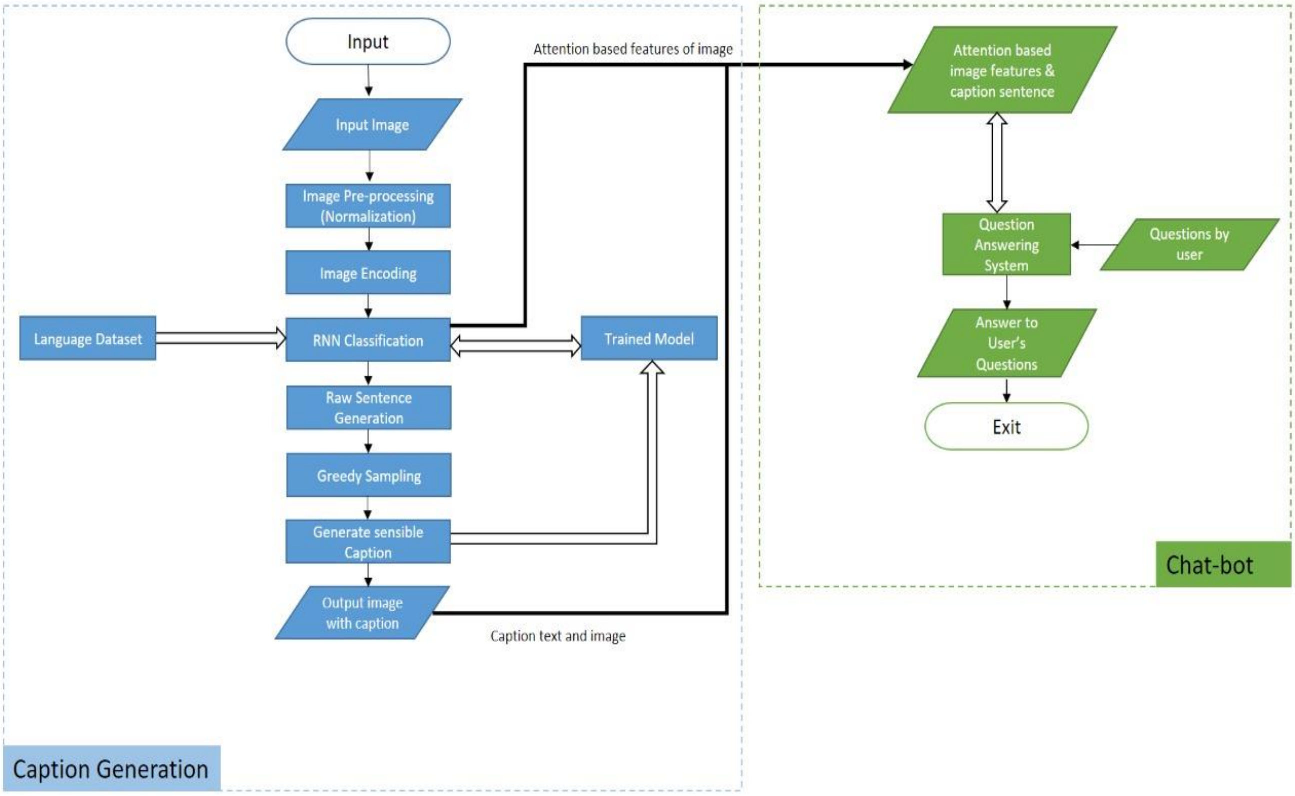

Input
Input Image
Image Pre-processing (Normalization)
Image Encoding
Language Dataset
RNN Classification
Trained Model
Raw Sentence Generation
Greedy Sampling
Generate sensible Caption
Output image with caption
Attention based features of image
Caption text and image
Caption Generation
Attention based image features & caption sentence
Question Answering System
Questions by user
Answer to User's Questions
Exit
Chat-bot

Fig 1: Proposed system architecture

A. Image Pre-processing: Pre-processing is a common name for operations with images at the lowest level of abstraction - both input and output are intensity images. The aim of preprocessing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing.

B. Image Encoding: It is also known as "encoding method". Image encoding is used to prepare photos to be displayed in a way that most computers and software, as well as browsers, can support. This is often necessary because not all image viewing software is able to open certain image files.

C. RNN Classification: Recurrent Neural Networks Recurrent Neural Networks (RNN) are a type of Neural Network where the output from the previous step is fed as input to the current step.

D. Language Dataset: Here we have used SQuAD dataset. Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable.

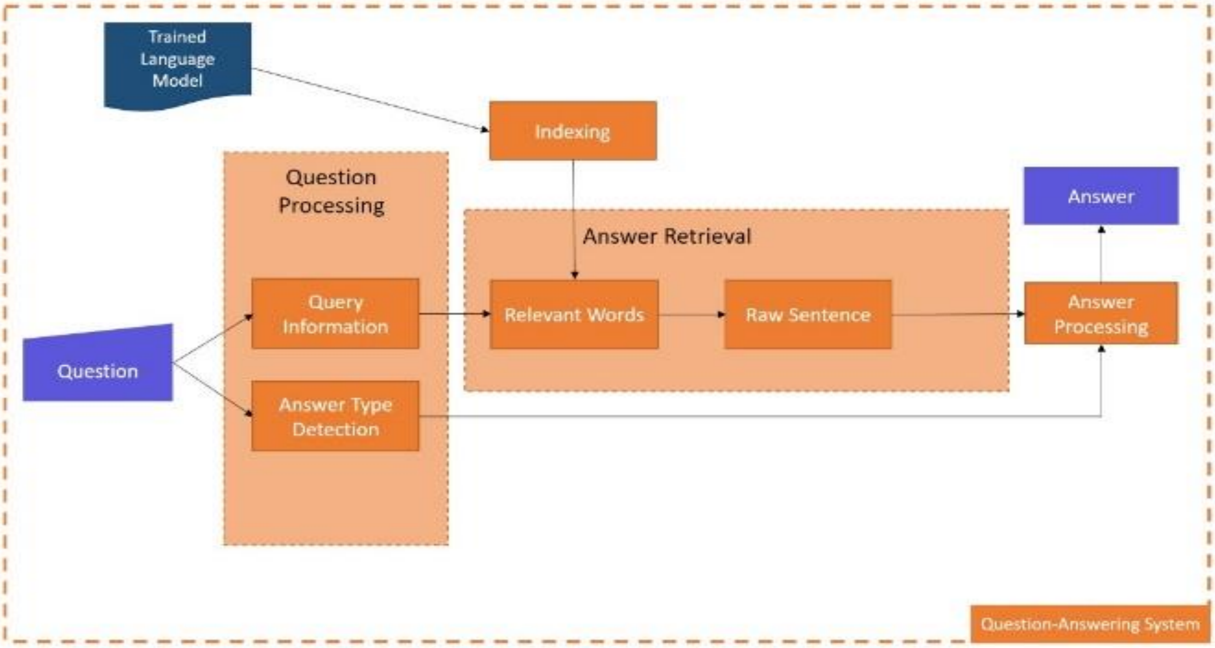


Fig 2: Chatbot Architecture

E. Information Retrieval-based QA: IR-based QA systems are sometimes called text-based and relies on unstructured corpus - huge amount of paragraphs on web sites such as news sites or Wikipedia. As can be seen from its name IR strategies are utilized to remove sections that can contain a response to given inquiry. The key phrases or keywords from a question which determine answer type make search query for search engine. The search engine returns the documents which are split into many passages. The last conceivable answer strings are browsed those entries and most fitted answer is chosen accordingly. A large portion of present day open-space QA frameworks are IR-based.

F. Question Processing: After question processing, some significant details extricated from an inquiry. Based on this information our task is to determine type of answer. It is called answer type recognition, or just question classification and rely on name entity recognition in most of the cases.

G. Document Processing: The subsequent stage is detailing of queries. For that we use query reformulation rules. The built inquiry is shipped off data recovery motor running dependent on various recorded archives. As a result, we get a set of documents which are ranked by relevance. The next step is retrieving units – passages, sentences or sections from a large set of documents. First we filter documents which do not contain the entities we got from answer type recognition phase. Secondly, we filter and rank other documents with using simple machine learning.

H. Answer Processing: The last step is an extraction of answer for an inquiry from the chose section or sentence. The main part of this work will focus on an answer processing. I will analyze different advanced sequential models based on embeddings of the given question and the selected passage.

4. Algorithms/ Techniques:

Convolutional Neural Network (CNN): CNN image classifications takes an input image, process it and classify it under certain categories (Eg- Dog, Cat, Tiger, Lion). Computers sees an input image as array of pixels and it depends on the image resolution. Based on the image resolution, it will see h x w x d(h = Height, w = Width, d = Dimension). Eg., An image of 6 x 6 x 3 array of matrix of RGB (3 refers to RGB values) and an image of 4 x 4 x 1 array of matrix of grayscale image. Technically, deep learning CNN models to train and test, each input image will pass it through a series of convolution layers with filters (Kernels), Pooling, fully connected layers (FC) and apply Softmax function to classify an object with probabilistic values between 0 and 1. Convolution of an image with different filters can perform operations such as edge detection, blur and sharpen by applying filters.

Recurrent Neural Network (RNN): Recurrent Neural Network is a generalization of feed forward neural network that has an internal memory. RNN is recurrent in nature as it performs the same function for every input of data while the output of the current input depends on the past one computation. After producing the output, it is copied and sent back into the recurrent network. For making a decision, it considers the current input and the output that it has learned from the previous input. Unlike feed forward neural networks, RNNs can use their internal state (memory) to process sequences of inputs. This makes them applicable to tasks such as unsegmented, connected handwriting recognition or speech recognition. In other neural networks, all the inputs are independent of each other. But in RNN, all the inputs are related to each other.

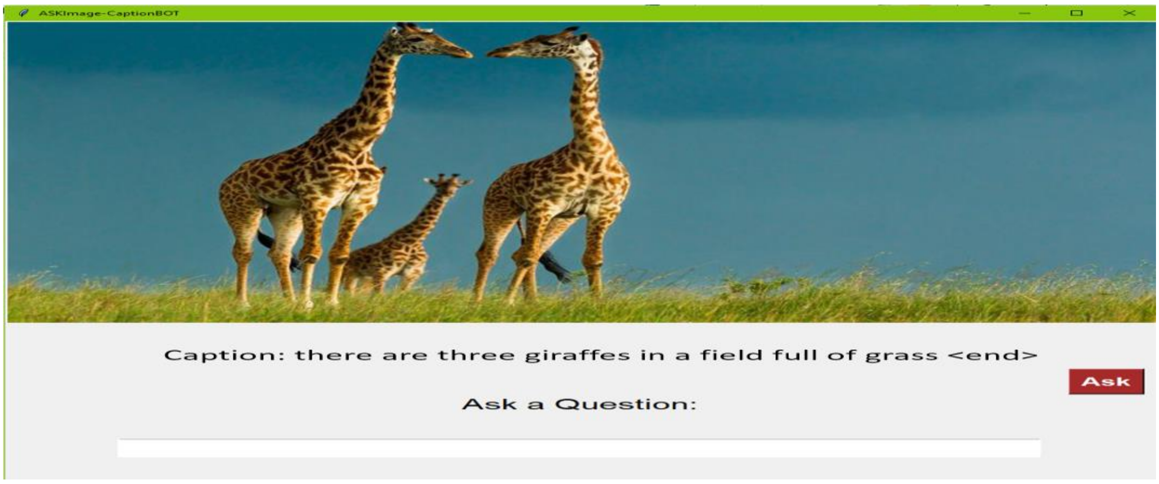
5. Dataset and Parameters

We have used COCO dataset for image captioning SQuAD dataset for question-answering system.

COCO Dataset (Image Captioning): COCO is a large-scale object detection, segmentation, and captioning dataset. The COCO dataset stands for **Common Objects in Context**, and is designed to represent a vast array of objects that we regularly encounter in everyday life. The COCO dataset is labeled, providing data to train supervised computer vision models that are able to identify the common objects in the dataset. Of course, these models are still far from perfect, so the COCO dataset provides a benchmark for evaluating the periodic improvement of these models through computer vision research.

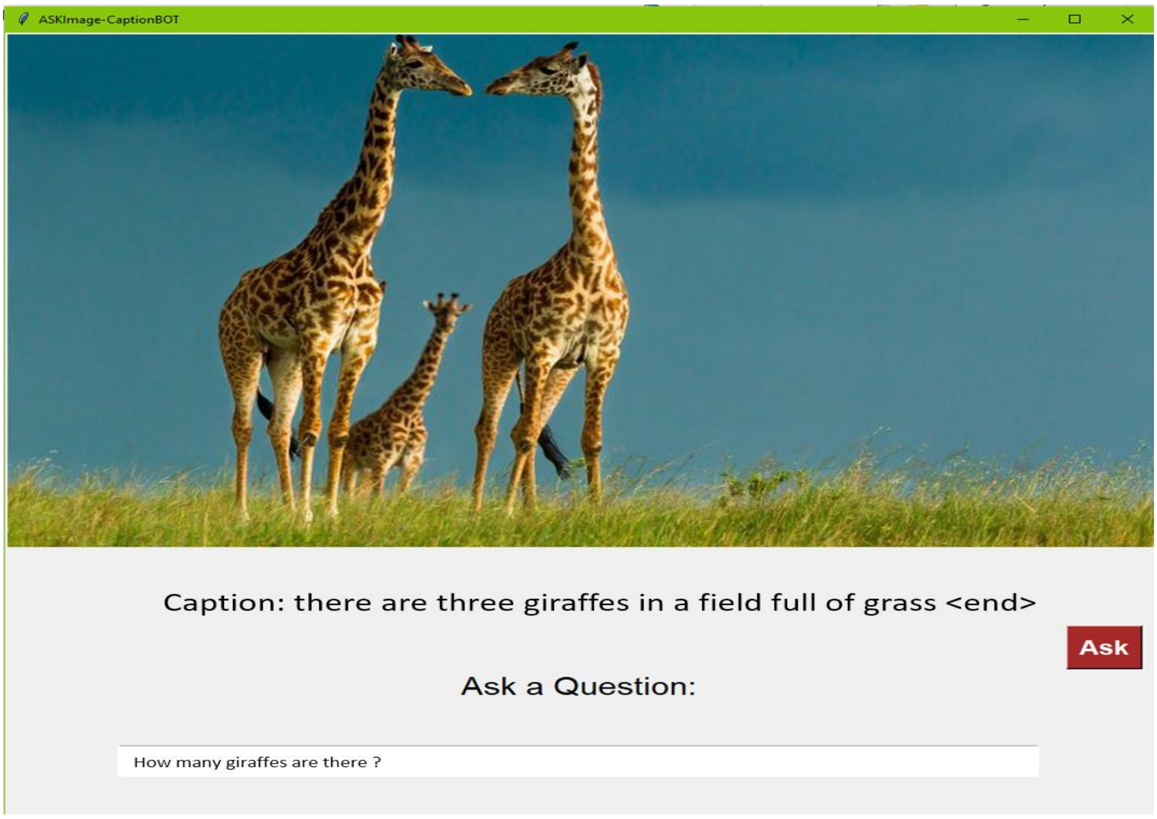
SQuAD Dataset (Question-Answering System): Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset, consisting of questions posed by crowd workers on a set of Wikipedia articles, where the answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. It combines 100,000 answerable questions with 50,000 unanswerable questions about the same paragraph from a set of Wikipedia articles. The unanswerable questions were written adversarially by crowd workers to look similar to answerable ones.

6. Implementation



Caption: there are three giraffes in a field full of grass <end>

Fig 3: Image Captioning

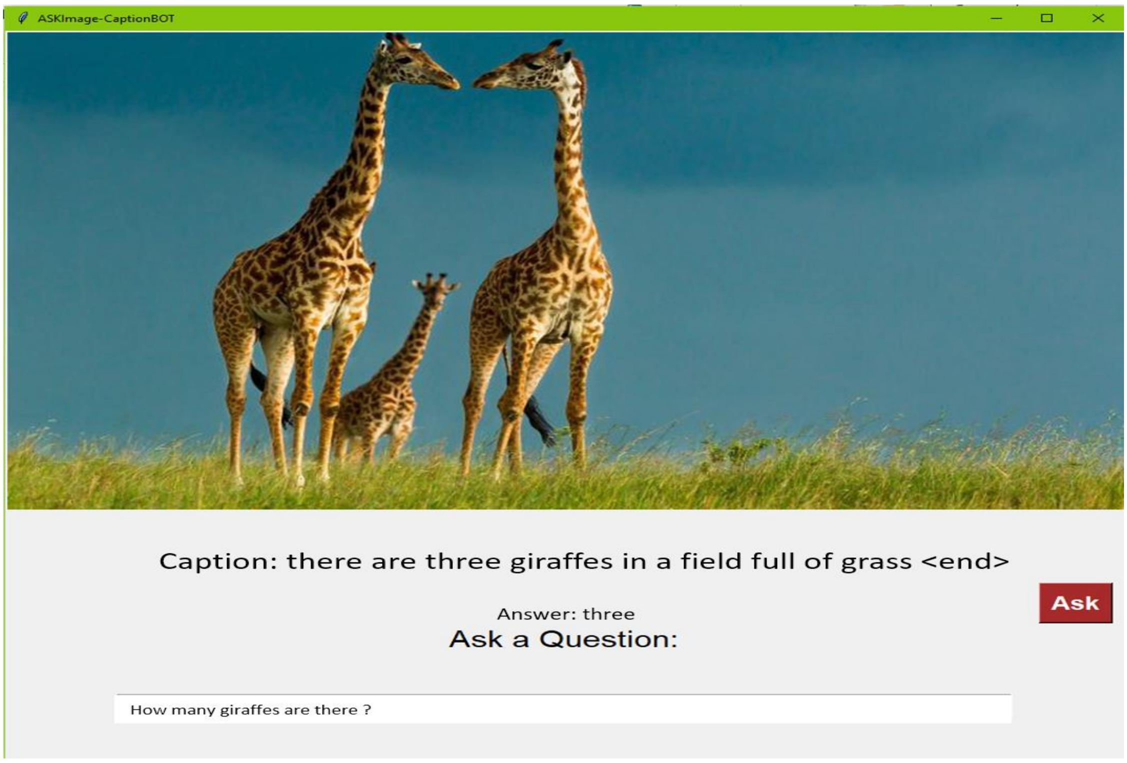


Caption: there are three giraffes in a field full of grass <end>

Ask a Question:

How many giraffes are there ?

Fig 4: Question-Answering System



Caption: there are three giraffes in a field full of grass <end>

Answer: three

Ask a Question:

How many giraffes are there ?

Fig 5: Question-Answering System

7. Future Scope

Although image caption can be applied to image retrieval, video caption, and video movement and the variety of image caption systems are available today, experimental results show that this task still has better performance systems and improvement. It mainly faces the following three challenges: first, how to generate complete natural language sentences like a human being; second, how to make the generated sentence grammatically correct; and third, how to make the caption semantics as clear as possible and consistent with the given image content. For future work, we propose the following four possible improvements:

1) An image is often rich in content. The model should be able to generate description sentences corresponding to multiple main objects for images with multiple target objects, instead of just describing a single target object.
2) For corpus description languages of different languages, a general image description system capable of handling multiple languages should be developed.
3) Evaluating the result of natural language generation systems is a difficult problem. The best way to evaluate the quality of automatically generated texts is subjective assessment by linguists, which is hard to achieve. In order to improve system performance, the evaluation indicators should be optimized to make them more in line with human experts' assessments.
4) A very real problem is the speed of training, testing, and generating sentences for the model should be optimized to improve performance.

8. Conclusion

In this paper, the study of "**ASK IMAGE: A chatbot which answers questions on image captions**" is presented. The different techniques such as Artificial Neural Network, Convolutional Neural Network in the domain of Deep Learning and Artificial Intelligence along with in-built APIs is explained. The comparative study of various techniques available is presented in this report. We must understand that the images used for testing must be semantically related to those used for training the model. For example, if we train our model on the images of cats, dogs, etc. we must not test it on images of air planes, waterfalls, etc. This is an example where the distribution of the train and test sets will be very different and in such cases no Machine Learning model in the world will give good performance. The applications of this domain are identified and presented.

REFERENCES

[1] Amir Vakili and Azadeh Shakery, "Enriching Conversation Context in Retrieval-based Chatbots" University of Tehran, 6 Nov 2019.

[2] Shuang Bai, School of Electronic and Information Engineering, Beijing Jiaotong University, No.3 Shang Yuan Cun, Hai Dian District, Beijing, China and Shan An, Beijing Jingdong Shangke Information Technology Co., Ltd, Beijing, China; "A Survey on Automatic Image Caption Generation"; May 2018.

[3] Vibhor Sharma, Monika Goyal and Drishti Malik, "An Intelligent Behaviour Shown by Chatbot System", International Journal of New Technology and Research (IJNTR) ISSN:2454-4116, Volume-3, Issue-4, April 2017.

[4] Andrej Karpathy;"Connecting Images and natural language." A dissertation submitted to the Department of Computer Science and the Committee on Graduate Studies of Stanford University in partial fulfilment of the requirements for the Degree of Doctor of PHILOSOPHY; August 2016.

[5] Anjana Tiha, "Intelligent Chatbot using Deep Learning", UID : U00619942 University of Memphis Spring, 2018 Date: 04/26/2018

[6] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio; "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention"; arXiv:1502.03044v3 [cs.LG] 19 Apr 2016.