**A MINI PROJECT REPORT**

**On**

**IMAGE CAPTION GENERATOR USING CNN AND LSTM**

*Submitted by,*

S.Nakul Reddy 19J41A1252

J.Meghana 19J41A1228

D.Shashank 20J45A1202

K.Pandari 19J41A1235

*in partial fulfillment of the requirements for award of the degree*

*Of*

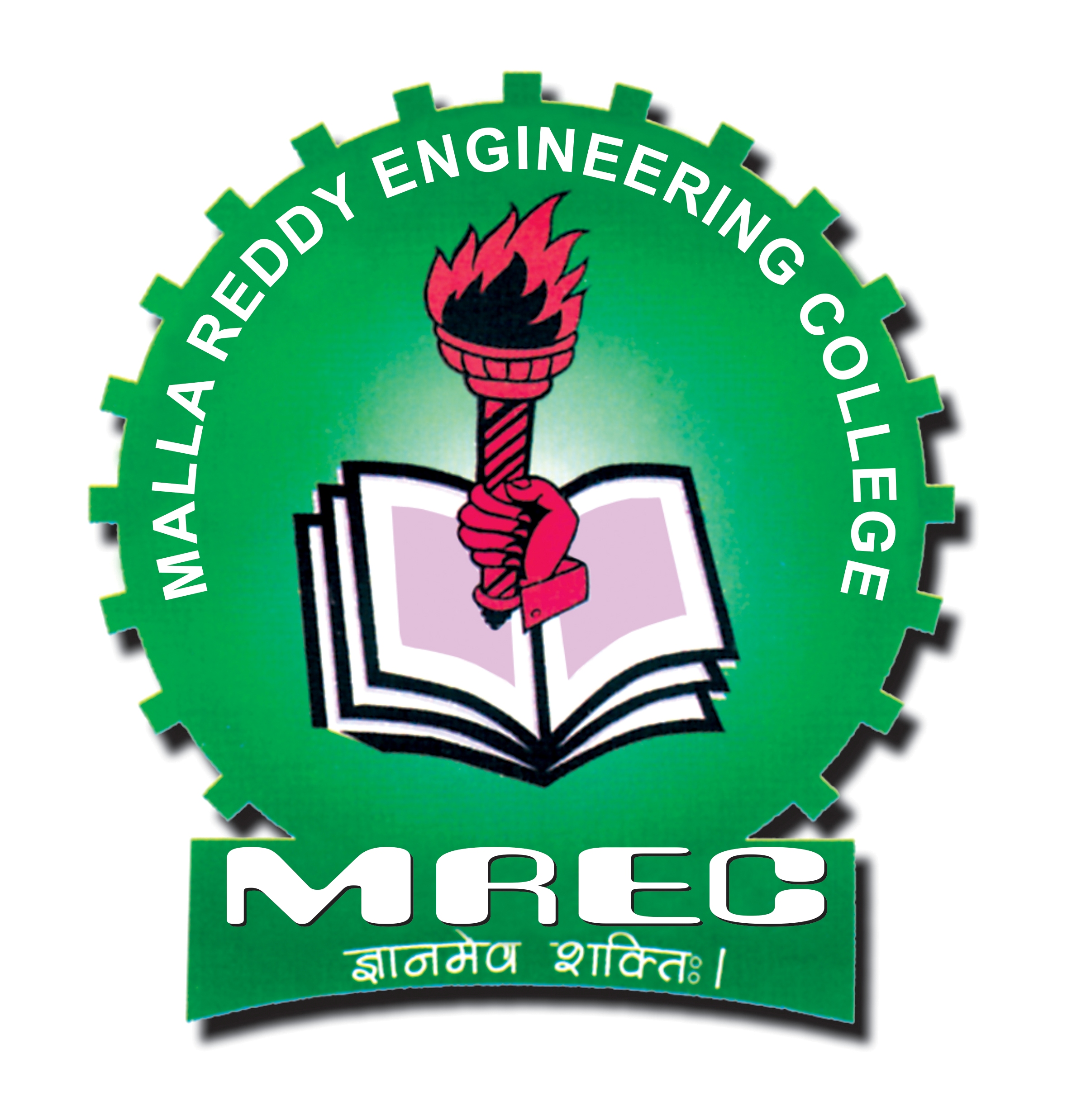
**BACHELOR OF TECHNOLOGY**

in

**INFORMATION TECHNOLOGY**

Under the Guidance of

**Mr.N.SATISH KUMAR**



**INFORMATION TECHNOLOGY**

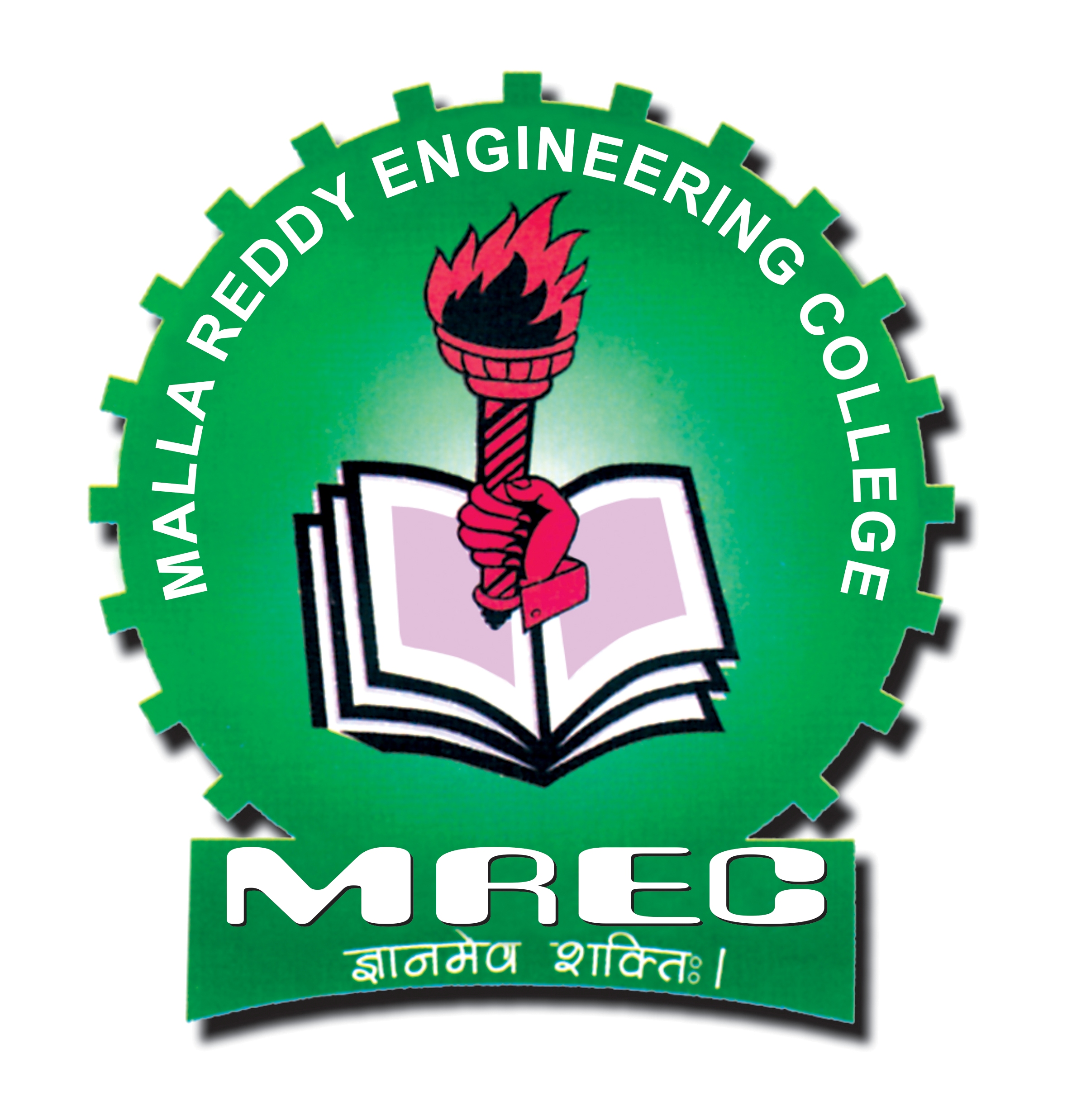
**MALLA REDDY ENGINEERING COLLEGE**

(An UGC Autonomous Institution, Approved by AICTE, New Delhi & Affiliated to JNTUH, Hyderabad) Maisammaguda, Secunderabad, Telangana, India 500100

**NOVEMBER-2022**

**MALLA REDDY ENGINEERING COLLEGE**

Maisammaguda, Secunderabad, Telangana, India 500100



**BONAFIDE CERTIFICATE**

This is to certify that this major project work entitled “IMAGE CAPTION GENERATOR USING CNN AND LSTM”, submitted by S.Nakul Reddy (19J41A1252), J.Meghana (19J41A1228), D.Shashank (20J45A1202), K.Pandari (19J41A1235) to Malla Reddy Engineering College affiliated to JNTUH, Hyderabad in partial fulfillment for the award of Bachelor of Technology in Name of the Programme is a bonafide record of project work carried out under my/our supervision during the academic year 2022 – 2023 and that this work has not been submitted elsewhere for a degree.

**SUPERVISOR** **HOD**

Department of Information Technology Department ofInformation Technology

Malla Reddy Engineering College Malla Reddy Engineering College

Secunderabad, 500 100 Secunderabad, 500 100

**Submitted for Major Project viva-voce examination held on \_\_\_\_\_\_\_\_\_**

**INTERNAL EXAMINER EXTERNAL EXAMINER**

i

**IMAGE CAPTION GENERATOR USING CNN AND LSTM**

ii

**MALLA REDDY ENGINEERING COLLEGE**

Maisammaguda, Secunderabad, Telangana, India 500100

**ACKNOWLEDGEMENT**

We are extremely thankful to our beloved Chairman and Founder of Malla Reddy Group of Institutions, **Sri.Ch.Malla Reddy**, for providing necessary infrastructure facilities for completing project work successfully.

We express our sincere thanks to our Principal, **Dr.A.Ramaswamy Reddy**, who took keen interest and encouraged us in every effort during the project work.

We express our heartfelt thanks to **Dr. Deena Babu Mandru**,Professor and Head, Department of Information Technology, for his kind attention and valuable guidance throughout the project work.

We are thankful to our Project Coordinator **Mr.N.Satish Kumar**, Assistant Professor, Department of Information Technology for his cooperation during the project work.

We are extremely thankful to our Project Guide **Mr.N.Satish Kumar** , Assistant Professor, for his constant guidance and support to complete the project work.

We also thank all the teaching and non-teaching staff of Department for their cooperation during the project work.

**S.Nakul Reddy 19J41A1252**

**J.Meghana 19J41A1228**

**D.Shashank 20J45A1202**

**K.Pandari 19J41A1235**

iii

**ABSTRACT**

Image captioning is one of the most needed requirements of today’s world. Moreover, there are some inbuilt applications that generate and provide a caption for a certain image, all these things are done with the help of deep neural network models. The process of generating a description of an image or the process by which we get to know moto of image is called image captioning. Using CNN it detects the important objects, their

attributes and also relation among them in respective image. It generates syntactically and semantically correct sentences. In this project, we present a CNN and LSTM Model to describe images and generate captions using computer vision and machine translation. This project aims to detect different and important objects found in an image, recognize the relationships between those objects and predicting the sequence of sentence using LSTM. The dataset used is MSCOCO and the programming language used is Python3, to demonstrate the caption for image. This model generates a decent description utilizing the trained data. To extract features from images we need encoder, we use CNN as encoder. To decode the description of image generated we use LSTM. This project will also elaborate on the functions and structure of the various Neural networks involved. Generating image captions is an important aspect of Computer Vision and Natural language processing.

*Keywords: CNN, LSTM, Natural language processing, Image captioning*

iv

**TABLE OF CONTENTS**

**ABSTRACT**  iv

**LIST OF FIGURES** vii

**LIST OF SYMBOLS AND ABBREVIATIONS** viii

**Chapter Description Page No**.

**1 INTRODUCTION**  1

**2 LITERATURE REVEIW**  4

**3 PROPOSED METHODOLOGY 7**

3.1. System Architecture 7

3.2 Algorithms 9

3.2.1 Convolution Neural Network 9

3.2.2 Long Short Term Memory 13

3.3. Overall Structure 16

**4 RESULTS AND DISCUSSION** 17

4.1. Requirement specification document 17

4.1.1. Functional requirements 17

4.1.2. Non-Functional requirements 17

4.2. Software requirements 18

4.2.1 Python packages 18

4.3. Results 18

v

**5 CONCLUSION 22**

5.1. Conclusion 22

**REFERENCES 23**

vi

**LIST OF FIGURES**

**FIGURE NO. TITLE PAGE NO.**

3.1 System Architecture 8

3.2 Image matrix multiples kernel or filter matrix 9

3.3 Matrix representation of image 10

3.4 Fully connected layer 11

3.5 Layers in CNN 11

3.6 Structure of CNN 12

3.7 Structure of LSTM 16

3.8 Overall structure of image caption generator 16

4.1 Output Screen 18

4.2 Input Image Uploaded 19

4.3 Running Existing Technique To Get Features 20

4.4 Window Showing The Caption Generator Of An 20

Input Image

4.5 Comparison graph 21

vii

**LIST OF SYMBOLS AND ABBREVIATIONS**

CNN - Convolutional Neural Network

LSTM - Long Short Term Memory

RNN - Recurrent Neural Network

AIA - Recurrent Neural Network

NLP - Natural Language Processing

LDA - Latent Dirichlet Allocation

SRS - Software Requirement Specification

viii

**CHAPTER 1**

**INTRODUCTION**

SOCIAL media are online communication channels dedicated to community- based input, interaction, content sharing, and collaboration. These media give the users the opportunity to share their content such as text, video, and images. Users usually accompany the content they post with text such as comments or hashtags. This alternative text (comment, hashtags, etc.) provides valuable information about the user posts and other information. Preece et al. construct a Sentinel platform that can enhance social media data in order to understand different situations they based also in YouTube video comments. Sagduyu et al. present a novel system that can present large-scale synthetic data from social media. In their system, they use textual content (hashtags and hyperlinks in tweets) to produce topics and train the n-gram model. The users in several of those media, e.g. Twitter, Instagram, and Facebook, use hashtags to annotate the digital content they upload. Hashtags are, usually, words or non spaced phrases preceded by the symbol that allow creators/content contributors to apply tagging that makes it easier for other users to locate their posts. A great portion of the digital content shared on social media platforms consists of images and short videos. Thus, effective retrieval of images from social media and the web, in general, becomes harder and more challenging day by day. Contemporary search engines are basically based on text descriptions to retrieve images; however, inaccurate text descriptions and the plethora of non textually annotated images led to extended research for content-based image retrieval techniques. The main problem of the content-based image retrieval is the so-called semantic gap: content-based retrieval is associated with low-level features while humans use high-level concepts for their search. To overcome this problem, automatic image annotation (AIA) methods were developed, that is, processes by which computing systems automatically

assign metadata in the form of captions or keywords to images. Among the AIA methods,

1

those based on the learning by example paradigm are probably the most common one. A small set of manually annotated training images are used to train models, which learn the correlation between image features and textual words (high-level concepts) and then allow automatic annotation of other (unseen) images. Obviously, good training examples, i.e., representative and accurate pairs of images and related tags are vital in this case.

Social media, and especially the Instagram, provide a rich source of image–tag pairs. Mining the right ones, automatically or semi automatically, so as to be used as training examples is extremely important.We have to consider, however, that, in many cases, hashtags that accompany images in social media are not related with the image’s content but serve several other purposes such as the expression of user’s emotional state, the increase in user’s clicks and find ability, and the beginning of a new communication or discussion. In our previous research, we have shown that the percentage of the Instagram hashtags that describe the visual content of the image they are associated with does not exceed 25%. We have also noticed that many Instagram hashtags are used across images that have nothing in common, just for search-ability enhancement. We named those hashtags as stop hashtags. Thus, filtering the Instagram hashtags in terms of the visual content of the image they accompany is required. Hyperlink-induced topic search (HITS) is a ranking algorithm than we could use to filter Instagram hashtags and locate the most relevant. The purpose of the HITS algorithm, developed by Jon Klein berg, is to rate web pages. The basic idea is that a web page can provide information about a topic and also relevant links for a topic. Thus, web pages belong to two groups: pages that provide good information about a topic (“authoritative”) and those that give to the user good links about a topic (“hubs”). The HITS algorithm gives to each web page both a hub and an authoritative value. We have started experimenting with the HITS algorithm for mining informative Instagram hashtags in one of our previous works and we extend this paper here by considering the application of the HITS algorithm in a real crowd tagging environment facilitated by the Figure-eight, formerly known as Crowd flower, crowd sourcing platform. In addition, we have increased the number of annotations per image to 500, we formed the bipartite graphs for all images, and we calculated the performance of annotators across all those images. Moreover, Folk Rank is used as a baseline to evaluate the performance of the proposed method.

2

The contents of a picture are automatically created in Artificial Intelligence (AI), which combines computer vision and natural language processing (NLP) (Natural Language Processing). It is developed a regenerative neuronal model. Computer vision and machine translation are required. This model is used to produce natural-sounding phrases that describe the picture. Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) are used in this model (RNN). The CNN is used to extract features from images, while the RNN is used to generate sentences. The model has been trained in such a manner that when an input image is provided to it, it creates captions that almost accurately describe the image. On various datasets, the model's accuracy, smoothness, and command of language learned from picture descriptions are assessed. These tests reveal that the model typically provides correct descriptions of the input image.

**PROJECT FRAMEWORK**

The entire research study has been subdivided into five parts. The first part is the introduction. This section provides the background of the study, research aim, objectives, and research significance. The second part is the literature review. This section provides the various processes and strategies of the entire study from existing research notes. The third part is the work methodology. This part has mainly described the methods for successfully conducting fundamental research. The fourth part is the finding and analysis, and the last part is the conclusion of the fundamental research

1. Introduction
2. Literature Survey
3. Proposed Methodology
4. Results and Discussion
5. Conclusion

3

**CHAPTER 2**

**LITERATURE REVIEW**

we will talk about the experimental results were carried out by MSCOCO dataset . For encoder/decoder framework , they have added a feature called guiding network in their proposed work . The method that called guiding network , mainly deals to learn the vector by a neural network v=g(A) where A is the set of annotation vectors . Generating natural language descriptions from visual data is an important problem . it has long been studied in computer vision . Hence, this had led to complex systems consists of visual primitive recognizer combined with a structured formal language like And-Or Graphs or logic systems .Recently, the problem of still image description with natural text has gained a huge interest.

**Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation**

Automatic Image Annotation (AIA) is the process of assigning tags to digital images without the intervention of humans. Most of the modern automatic image annotation methods are based on the learning by example paradigm. In those methods building the training examples, that is, pairs of images and related tags, is the first critical step. We have shown in our previous studies that hashtags accompanying images in social media and especially the Instagram provide a reach source for creating training sets for AIA. However, we concluded that only 20% of the Instagram hashtags describe the actual content of the image they accompany, thus, a series of filtering steps need to apply in order to identify the appropriate hashtags. In this paper we apply topic modelling with Latent Dirichlet Allocation (LDA) on Instagram hashtags in order to predict the subject of the related images.Since a topic is composed by a set of related terms, the identification of the visual topic of an Instagram image, through the proposed method, provides a plausible set of tags to be used in the context of training AIA methods.

4

**Crowd sourcing for multiple-choice answering**

We leverage crowd wisdom for multiple-choice question answering, and employ lightweight machine learning techniques to improve the aggregation accuracy of crowd sourced answers to these questions. In order to develop more effective aggregation methods and evaluate them empirically, we developed and deployed a crowd sourced system for playing the “Who wants to be a millionaire?” quiz show. Analyzing our data (which consist of more than 200,000 answers), we find that by just going with the most selected answer in the aggregation, we can answer over 90% of the questions correctly, but the success rate of this technique plunges to 60% for the later/harder questions in the quiz show. To improve the success rates of these later/harder questions, we investigate novel weighted aggregation schemes for aggregating the answers obtained from the crowd. By using weights optimized for reliability of participants (derived from the participants’ confidence), we show that we can pull up the accuracy rate for the harder questions by 15%, and to overall 95% average accuracy. Our results provide a good case for the benefits of applying machine learning techniques for building more accurate crowd sourced question answering system

**Validity and reliability of naturalistic driving scene categorization judgments from crowd sourcing**

A common challenge with processing naturalistic driving data is that humans may need to categorize great volumes of recorded visual information. By means of the online platform Crowd Flower, we investigated the potential of crowd sourcing to categorize driving scene features (i.e., presence of other road users, straight road segments, etc.) at greater scale than a single person or a small team of researchers would be capable off.

Validity and reliability were examined, both with and without embedding researcher generated control questions via the Crowd Flower mechanism known as Gold Test Questions (GTQs). By employing GTQs, we found significantly more valid (accurate) and reliable (consistent) identification of driving scene items from external workers.

5

Specifically, at a small scale Crowd Flower Job of 48 three-second video segments, an accuracy (i.e., relative to the ratings of a confederate researcher) of 91% on items was found with GTQs compared to 78% without. A difference in bias was found, where without GTQs, external workers returned more false positives than with GTQs. At a larger scale Crowd Flower Job making exclusive use of GTQs, 12,862 three-second video segments were released for annotation. Infeasible (and self-defeating) to check the accuracy of each at this scale, a random subset of 1012 categorizations was validated and returned similar levels of accuracy (95%).

**A survey and analysis on Automatic Image Annotation**

In recent years, image annotation has attracted extensive attention due to the explosive growth of image data. With the capability of describing images at the semantic level, image annotation has many applications not only in image analysis and understanding but also in some relative disciplines, such as urban management and biomedical engineering. Because of the inherent weaknesses of manual image annotation, Automatic Image Annotation (AIA) has been raised since the late 1990s. In this paper, a deep review of state-of-the-art AIA methods is presented by synthesizing 138 literature published during the past two decades.

Authors classified AIA methods into five categories:

1. Generative model-based image annotation.

2) Nearest neighbor-based image annotation.

3) Discriminative model-based image annotation.

4) Tag completion-based image annotation.

5) Deep Learning-based image annotation.

Comparisons of the five types of AIA methods are made on the basis of the underlying idea, main contribution, model framework, computational complexity, computation time, and annotation accuracy.

6

**CHAPTER 3**

**PROPOSED METHODOLOGY**

Here We use CNN and LSTM to achieve our goal (image caption generator) we start from what is CNN and how can benefit from it in our problem ? Convolutional Neural Network is an artificial deep learning neural network. It is used for image classifications , computer vision ,image recognition and Object detection. CNN image classifications takes an input image, process it and classify it under certain categories (Eg., Dog, Cat,etc). 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. Secondly , what is LSTM ? 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. we merged this two models in one model called a CNN-RNN model.in general Our approach draws on the success of the top-down image generation models listed above. We use a deep convolutional neural network to extract the visual image features and Semantic features are extracted from the semantic tagging model. Visual features from CNN and semantic features from tagging model are concatenated and feed as the input to a Long-Short-Term Memory (LSTM) network, which then generates captions

**3.1. SYSTEM ARCHITECTURE**

The proposed system generates a caption for the image given as input. This is possible by training the deep learning model with the dataset. To increase the quality of

captions,we can do the cleaning of the dataset by removing unnecessary things and punctuation's. We can use semantic feature extraction and making vocabulary.

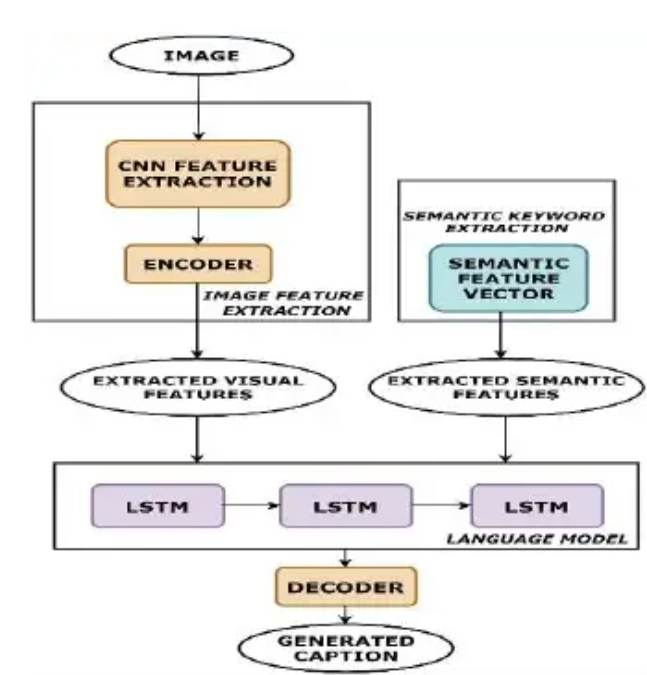
7

Using glove embedding can increase the quality of captions generated. By using an Encoder-Decoder model we can limit the length of the generated captions. So that our model will generate captions of that fixed length for a given input image.This model is mainly divided into three modules:

1. Image recognition model - It is used for the extraction of features for a given image.

**2.** Semantic feature extraction model - It extracts the specific keyword in the dataset which improves the quality of captions.

**3.** Language model - It will generate captions based on the features of the given image.



**Figure 3.1 : System Architecture**

8

**3.2. ALGORITHMS**

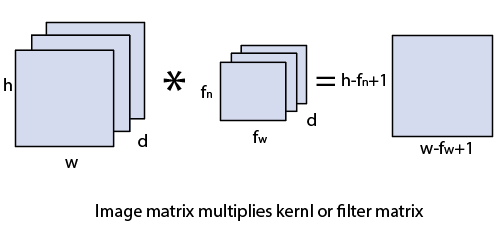
It involves 2 deep learning algorithms which are used for extracting key features and generating caption for the image

**3.2.1. CONVOLUTIONAL NEURAL NETWORK**

It involves extracting the features from the image like Color,Texture and patterns, Shape and objects, Position, etc... Using Convolutional neural network (CNN).Convolutional Neural Network (CNN) is used for identifying patterns and understanding them. It is a type of ANN. CNN's are a class of Deep Neural Networks that can detect and classify particular features from images and are widely used for analyzing visual images. Their main applications range from recognition of images, classification of images, analysis of medical images for disease identification, natural language processing, and computer vision.CNN has five different layers which will perform different tasks.

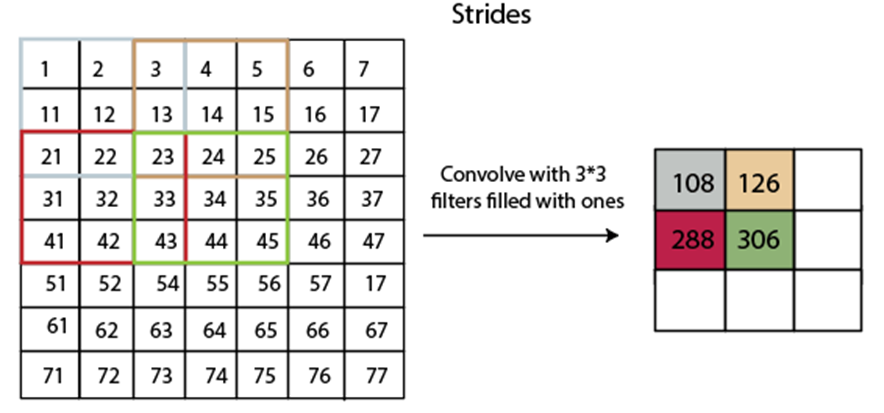
The five different layers in CNN are:

1. Input layer- Input layer contains the image data that is fed as input.
2. Convolutional layer - This layer is used for feature extraction for images that are fed as input to the layer.This layer performs the mathematical operation of convolution between a filter of a particular size MxM and the input image. The dot product is performed between the image and the filter. The output gives information about the input image such as the edges,corners, and patterns is called the Feature map. This output is given to other layers as input to get any other features if required.



**Figure 3.2: Image matrix multiples kernel or filter matrix**

**9**



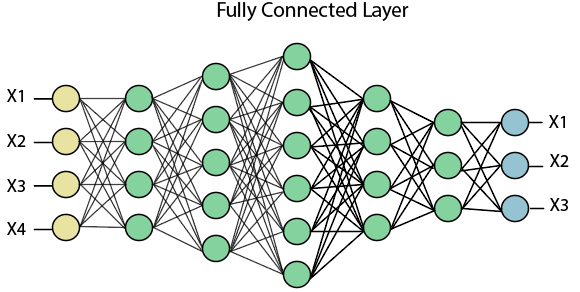
**Figure 3.3 : Matrix representation of image**

3.Activation function layer -This layer is used to learn and approximate any complex and continuous relationship in variables of a network. It is used to classify the data which can be forwarded in the network or data which can be left.

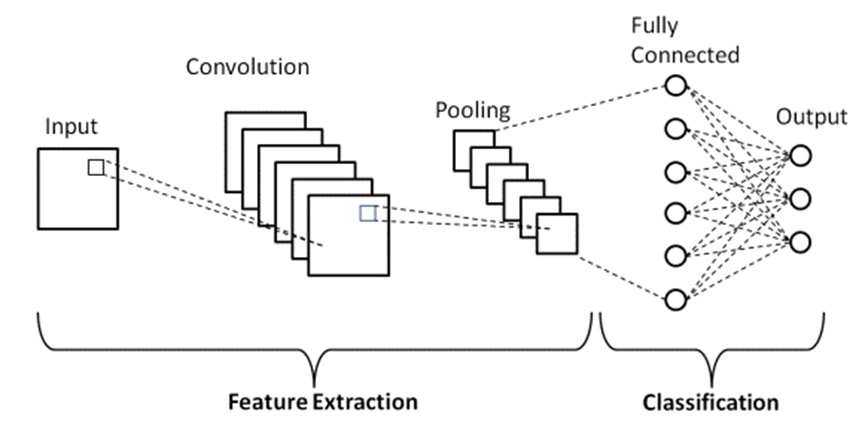
4. Pooling layer- The pooling layer is used for reducing computational costs by reducing the size of the feature map. This can be achieved by making only fewer connections between layers.

5.Fully connected layer - This layer consists of the neurons along with biases and weights. It is used to make a connection between the neurons of two different layers. It can be used for the classification of images after training.Output layer - holds the extracted feature values at the end of the process Inception v3 is a widely-used image recognition model that has been shown to attain greater than 78.1% accuracy on the Image Net dataset

10



**Figure 3.4 : Fully connected layer**



**Figure 3.5 : Layers in CNN**

11

Semantic Keyword extraction is the process of automatically identifying the terms that best describe the subject of a document, in this case, an image that is grammatically correct. The generated caption should consist of much more clear information on attributes of the image such as color. By using Semantic Keyword Extraction, the quality of caption swill be increased and the captions generated will be more grammatically correct.After getting the vocabulary or keywords, these are given as input to the perceptron network. This network is a multi-layered network with each layer designated with a unique function. Some layers like the dropout layer, soft max layer,and Dense layer are responsible for the identification of attributes while will help the model to generate meaningful captions. The steps involved are:

Step 1: The dataset will be cleaned by removing punctuation.

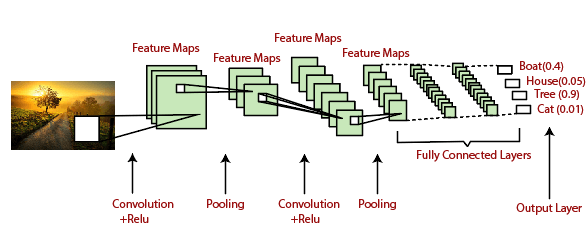
Step 2: Data loading and preprocessing.

Step 3: converting words to vectors

Step 4: Creating a vocabulary with unique words

Step 5: From the dictionary created in the previous step, the top 400 semantics will be selected.

Step 6:The vector that is created in the previous step is given to multi layered network



**Figure 3.6 : Structure of CNN**

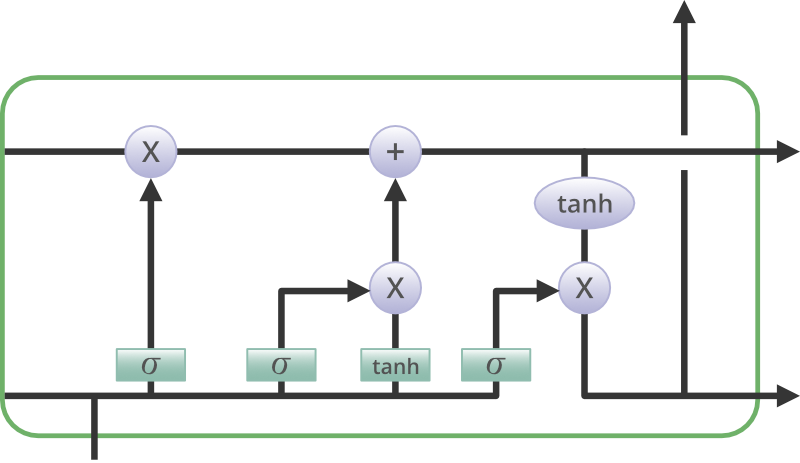
12

**3.2.2. LONG SHORT TERM MEMORY**

Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. LSTM was designed by Hochreiter & Schmidhuber. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give efficient performance. LSTM can by default retain the information for long period of time. It is used for processing, predicting and classifying on the basis of time series data.

**Structure of LSTM**

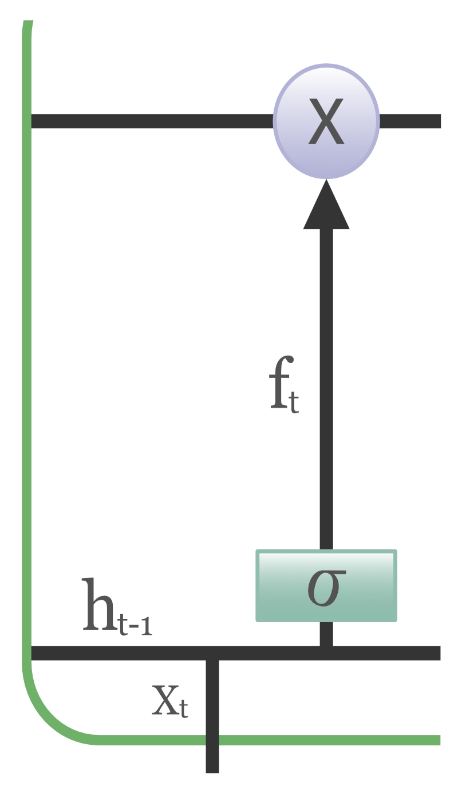
LSTM has a chain structure that contains four neural networks and different memory blocks called .



Information is retained by the cells and the memory manipulations are done by the**gates.** There are three gates –

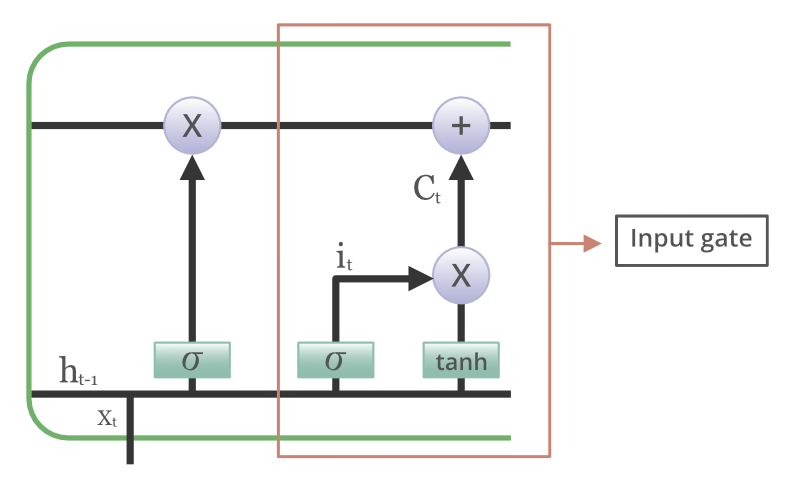
13

**Forget Gate:** The information that no longer useful in the cell state is removed with the forget gate. Two inputs *x\_t* (input at the particular time) and *h\_t-1* (previous cell output) are fed to the gate and multiplied with weight matrices followed by the addition of bias. The resultant is passed through an activation function which gives a binary output. If for a particular cell state the output is 0, the piece of information is forgotten and for the output 1, the information is retained for the future use.

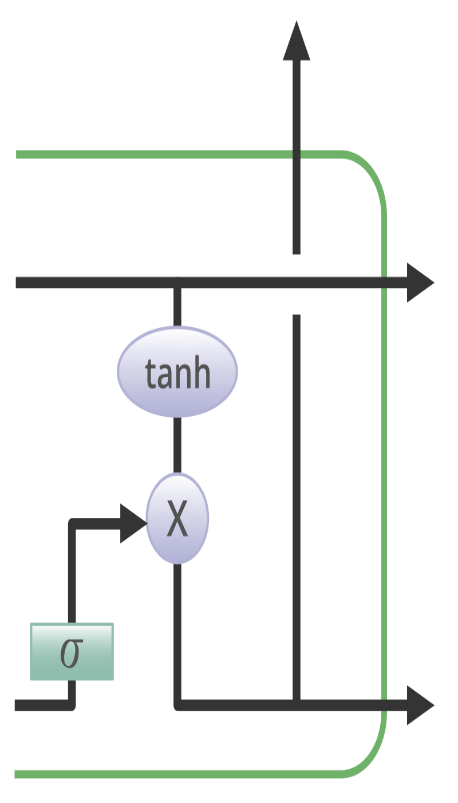


**Input gate:** Addition of useful information to the cell state is done by input gate. First, the information is regulated using the sigmoid function and filter the values to be remembered similar to the forget gate using inputs*h\_t-1* and *x\_t*. Then, a vector is created using*tanh*function that gives output from -1 to +1, which contains all the possible values from h\_t-1 and *x\_t*. At last, the values of the vector and the regulated values are multiplied to obtain the useful information.

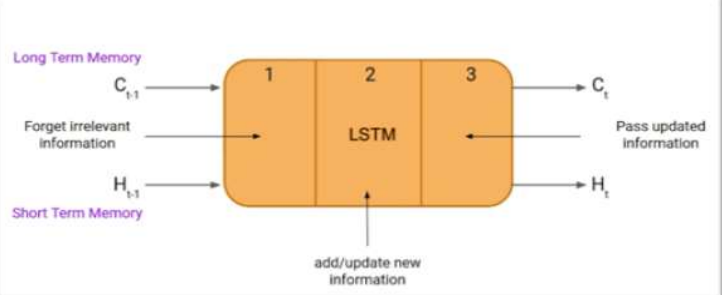
14



**Output gate:** The task of extracting useful information from the current cell state to be presented as an output is done by output gate. First, a vector is generated by applying tanh function on the cell. Then, the information is regulated using the sigmoid function and filter the values to be remembered using inputs*h\_t-1* and *x\_t*. Atlast, the values of the vector and the regulated values are multiplied to be sent as an output and input to the next cell.

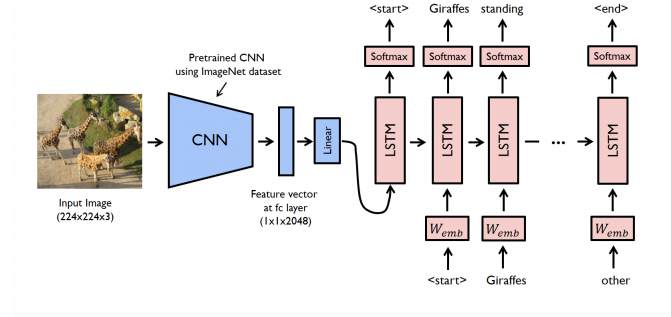


15



**Figure 3.7 : Structure of LSTM**

**3.3. OVERALL STRUCTURE**



**Figure 3.8 : Overall structure of image caption generator**

16

**CHAPTER 4**

**DISCUSSION AND RESULTS**

**4.1 REQUIREMENT SPECIFICATION DOCUMENT**

A software requirements specification (SRS) is a description of a software system to be developed. It lays out functional and nonfunctional requirements, and may include a set of use cases that describe user interactions that the software must provide.Software requirements specification establishes the basis for an agreement between customers and contractors or suppliers (in market-driven project, these roles may be played by the marketing and development divisions) on what the software product is to do as well as what it is not expected to do. Software requirements specification permits a rigorous assessment of requirements before design can begin and reduces later redesign

**4.1.1. FUNCRIONAL REQUIREMENTS**

* Load the dataset of images with different resolutions
* Encoding the image.
* Extracting the Features from the image.
* Recognizing the keywords in images.

**4.1.2. NON FUNCTIONAL REQUIREMENTS**

* Performance
* Response Time
* Throughput
* Serviceability
* Data Integrity

17

**4.2. SOFTWARE REQUIREMENTS**

Operating System : Windows/MAC

Programming Language : Python

Web Framework : tkinter

**4.2.1 PYTHON PACKAGES**

* Tensor flow
* Keras
* Numpy
* Tqdm

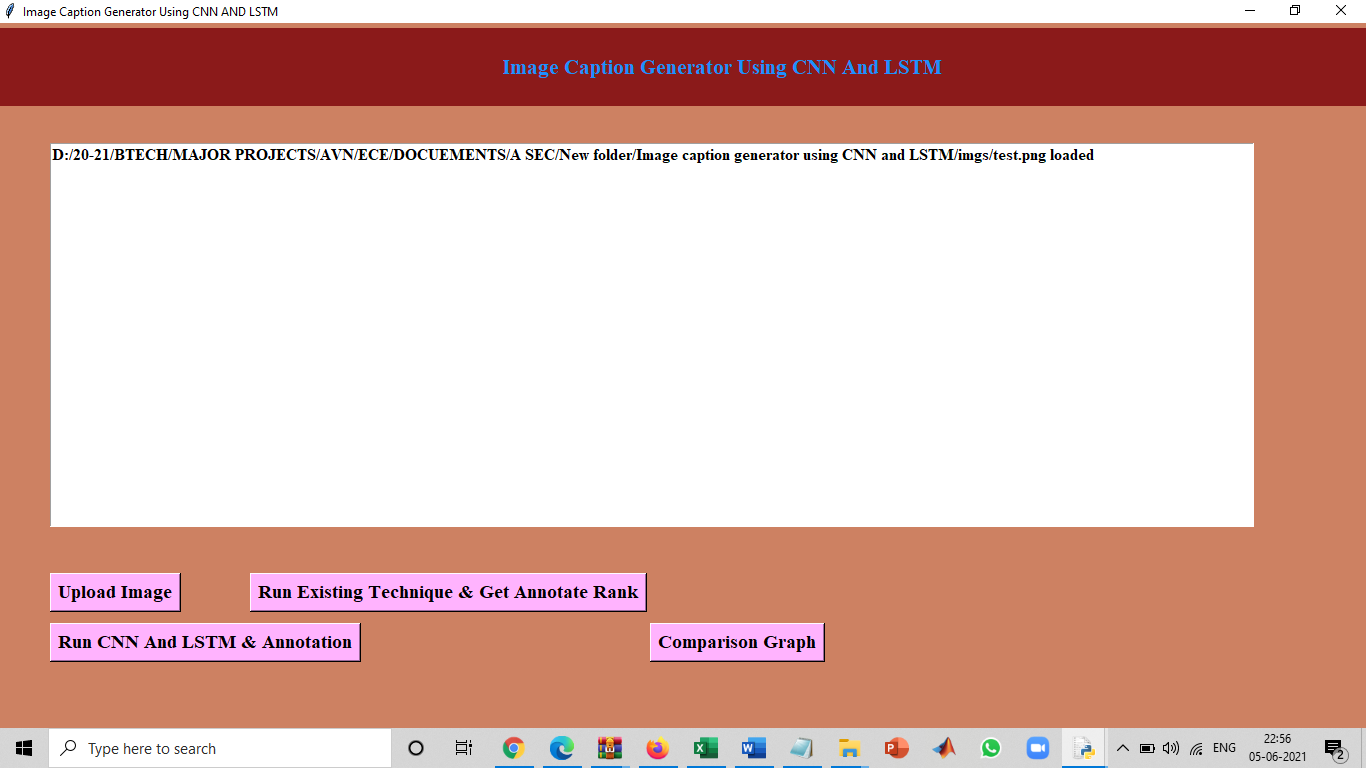
**4.3. RESULTS**



**Figure 4.1 : Output Screen**

18

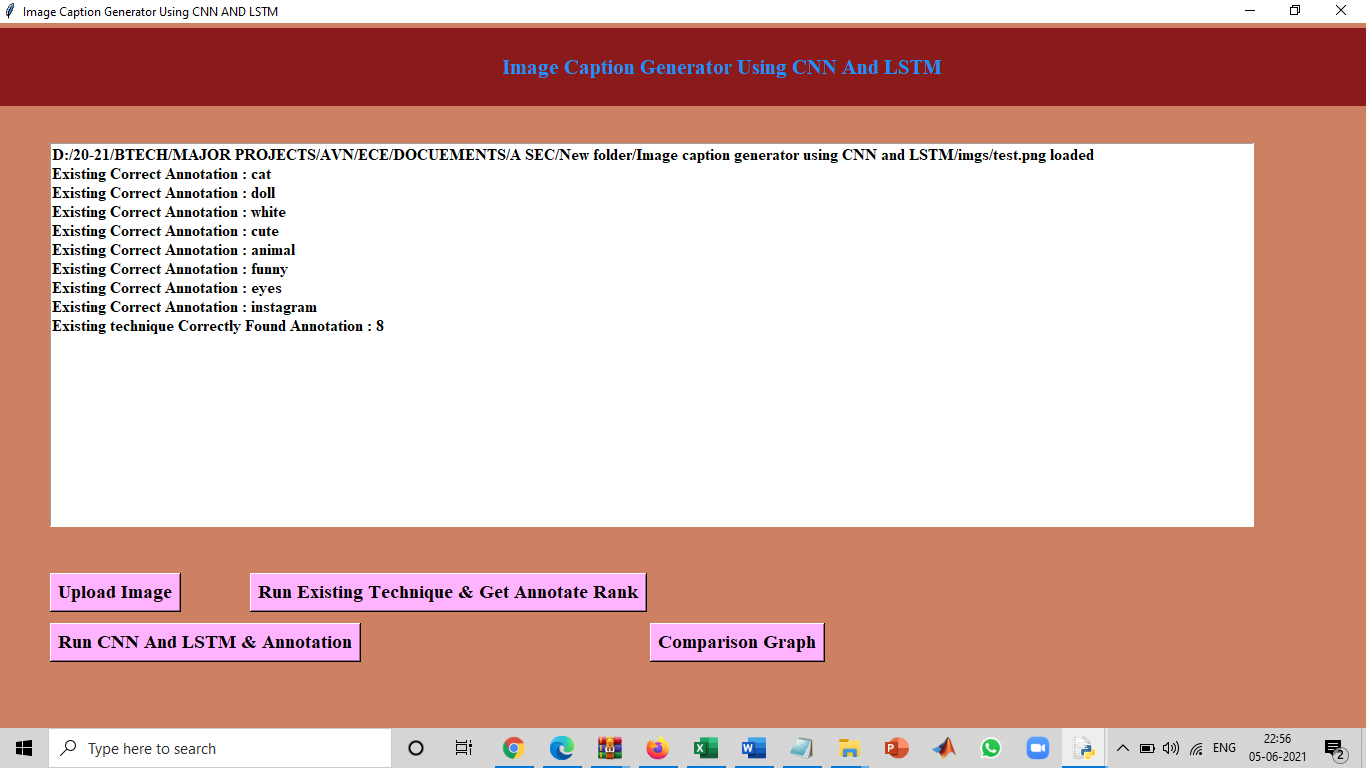
As we have used CNN and LSTM Algorithms and MSCOCO data set which created a well trained model that can generate accurate captions to the image. The generated caption is very accurate to the image given and can be used in various applications. Once we run the code it generates a output screen which is shown in the Fig 4.1.



**Figure 4.2 : Input Image Uploaded**

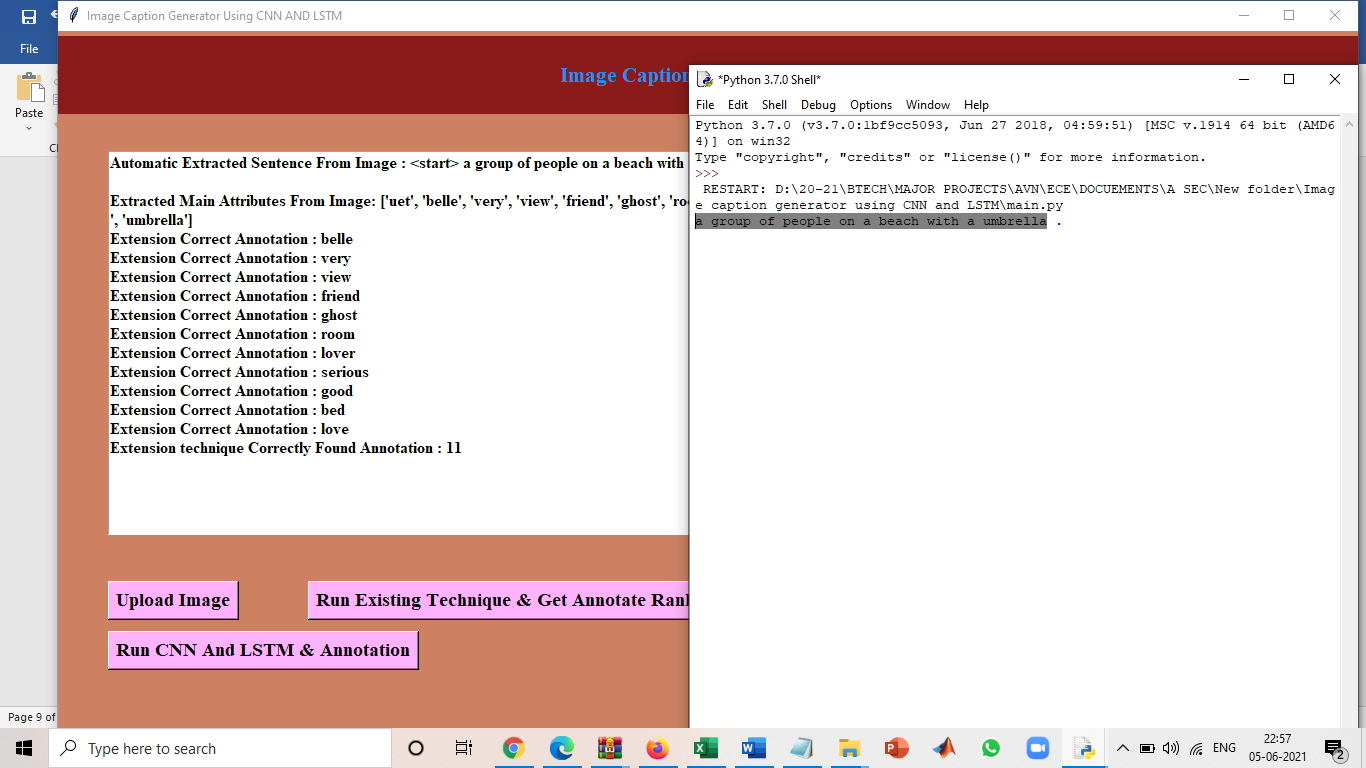
Next, we need to upload the image by using upload image button. Select the input image which we want to generate the caption. so image will be uploaded as the input shown in fig 4.2.

19



**Figure 4.3: Running Existing Technique To Get Features**

After the image is uploaded Run the existing technique and get annotate rank to the objects that detects for obtaining features from the input as shown in fig 4.3.

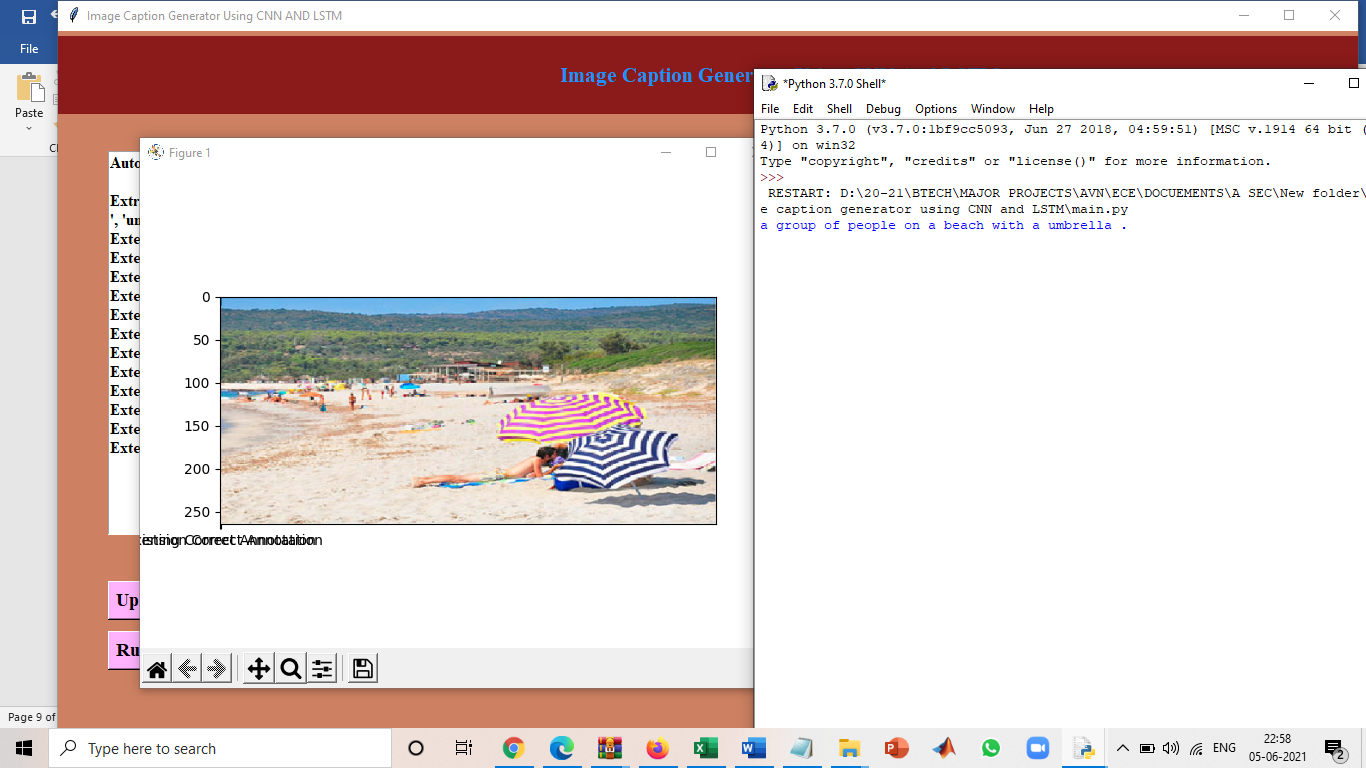


**Figure 4.4: Window Showing The Caption Generator Of An Input Image**

20

Then run the CNN and LSTM algorithms to generate the caption Hence window showing the caption generator of an input image as shown in fig 4.4.

And also there is comparison graph which shows our Image and the caption as well in the Fig 4.5



**Figure 4.5:Comparison graph**

21

**Chapter 5**

**CONCLUSION**

We conducted an extensive hyper parameter search over the CNN-LSTM model architecture, producing a best model that achieves results that are 3.3 BLEU-4 points and 3.8 CIDEr points behind the state-of-the-art, using a keep probability of 75% for dropout and two layers for our decoder LSTM network. A thorough quantitative and qualitative analysis of the generated metrics suggests that the model is able to sensibly caption a wide variety of images from the MSCOCO dataset. Partial errors tend to occur due to lack of attention to specific details in images (e.g. labeling a picture of elephants walking in an enclosure as ’elephants in a field’ due to being misdirected by trees in the background). This suggests that the attention-mechanisms explored in recent work may yield improvements to this task. Our main novel contribution to the field is exploring the effect of emitted words on hidden states in the LSTM that were previously treated as black boxes. We demonstrated that semantically-close emitted words (e.g. ’plate’ and ’bowl’) result in similar movements in hidden state despite different previous context and that divergences in hidden state occur only upon emission of semantically-far words (e.g. ’vase’ and ’food’).

In this advanced Python project, we have implemented a CNN-RNN model by building an image caption generator. Some key points to note are that our model depends on the data, so, it cannot predict the words that are out of its vocabulary. We used a small dataset consisting few images. For production-level models, we need to train on datasets larger than 100,000 images which can produce better accuracy models.

22

**REFERENCES**

[1] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. Every picture tells a story: Generating sentences from images. In Proceedings of the 11th European Conference on Computer Vision: Part IV, ECCV’10, pages 15–29, Berlin, Heidelberg, 2010. Springer-Verlag.

1. Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. Collective generation of natural image descriptions. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1, ACL ’12, pages 359–368, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.
2. Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Fifteenth Conference on Computational Natural Language Learning, CoNLL ’11, pages 220–228, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.
3. Xinlei Chen and C. Lawrence Zitnick. Learning a recurrent visual representation for image caption generation. CoRR, abs/1411.5654, 2014.
4. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L. Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). CoRR, abs/1412.6632, 2014.
5. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. CoRR, abs/1411.4555, 2014.
6. Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. CoRR, abs/1603.03925, 2016.

1. Ryan Kiros, Ruslan Salakhutdinov, and Richard S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. CoRR, abs/1411.2539, 2014.
2. Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. CoRR, abs/1411.4389, 2014.

23

1. Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. ¨ Neural Comput., 9(8):1735–1780, November 1997.
2. Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. CoRR, abs/1412.2306, 2014.
3. Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollar, ´ Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. From captions to visual concepts and back. CoRR, abs/1411.4952, 2014.
4. Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. Ask me anything: Dynamic memory networks for natural language processing. CoRR, abs/1506.07285, 2015.
5. Alex Graves. Generating sequences with recurrent neural networks. CoRR, abs/1308.0850, 2013.
6. Karol Gregor, Ivo Danihelka, Alex Graves, and Daan Wierstra. DRAW: A recurrent neural network for image generation. CoRR, abs/1502.04623, 2015.

24