BMS COLLEGE OF ENGINEERING

(Autonomous Institute, Affiliated to VTU)

Bull Temple Road, Basavanagudi, Bengaluru - 560019

A project report on

"Deep Learning Image Caption Generator"

Submitted in partial fulfilment of the requirements for the award of degree

BACHELOR OF ENGINEERING

IN

INFORMATION SCIENCE AND ENGINEERING

By

P Aishwarya Naidu (1BM16IS062)

Gehna Anand (1BM16IS034)

Satvik Vats (1BM16IS079)

Under the guidance of

Nalina V

Assistant Professor, Department of Information Science and Engineering

Department of Information Science and Engineering

2019-2020

BMS COLLEGE OF ENGINEERING

(Autonomous Institute, Affiliated to VTU)
Bull Temple Road, Basavanagudi,
Bengaluru – 560019

Department of Information Science and Engineering

C E R T I F I C A T E

This is to certify that the project entitled "**Deep Learning Image Caption Generator**" is a bona-fide work carried out by **P Aishwarya Naidu (1BM16IS062), Gehna Anand (1BM16IS034)** and **Satvik Vats (1BM16IS079)** in partial fulfillment for the award of degree of Bachelor of Engineering in **Information Science and Engineering** from **Visvesvaraya Technological University, Belgaum** during the year **2019-2020**. It is certified that all corrections/suggestions indicated for Internal Assessments have been incorporated in the report deposited in the departmental library. The project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Nalina V	**Dr. M Dakshayini**	**Dr. B V Ravishankar**
Assistant Professor	**Professor and HOD**	**Principal**

Examiners

Name of the Examiner	**Signature of the Examiner**
1.	
2.	

ABSTRACT

Computer vision has been an area of interest for engineers and scientists who had been spearheading in the field of artificial intelligence from the late 1960s as it was very essential to give the machines or robots the power of visualizing objects and activities around them like the human visual system. The ability to visualize 2-Dimensional images and extracting features from it can be utilised for developing various applications. The involvement of deep learning has been successful in bolstering the field of computer vision even further. The abundance of images in today's digital world and the amount of information contained in them has made it a very valuable and research worthy data item. A deep learning based image caption generator model is able to incorporate the areas of natural language and computer vision with deep learning to give a solution in which the machine is able to extract features from an image and then describe those features in a natural language. Thus, explaining the contents of the image in human readable format. This model has various applications ranging from social causes like being an aid to visually impaired to enhancing search experience of users over the web.

Thus, this report analyses the various state-of-the-art work in the field of image processing,computer vision and deep learning and presents a deep learning model that generates captions describing the images given as input to the system.

ACKNOWLEDGMENTS

We have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and our college. We would like to extend our sincere thanks to all of them.

We are highly indebted to Professor Nalina V. for her guidance and constant supervision as well as for providing necessary information regarding the project & also for her support in completing the project.

We would like to express our gratitude towards our parents for their kind cooperation and encouragement which helped us in completion of this project.

Our thanks and appreciations also go to our colleagues in developing the project and people who have willingly helped us out with their abilities.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

Chapter I

INTRODUCTION

I.I Overview

In this era of Information Technology when it is popularly known that data is the new oil, it is important to know why data has become the 'new oil' of the world. This digital age or the New Media Age has not only made us realize our dream of making the world a global village but has also made the long distances seem much shorter and removed barriers with every new discovery. This digital age also made our images 'digital' and in no time they became a large data set in the world. Data is the new oil because it is the fuel for the engine that drives the information economy, although use of data is no longer limited to economic purposes alone, it is now essential for almost every aspect of human civilization, from governance to business, travel and tourism to shopping, law enforcement to disaster management.

As energy is extracted from oil fuel in the same way nowadays information can be extracted from data. In that sense, Image data can be called as one of the fuels which has the highest 'calorific value', as the amount of information that can be extracted from it is enormous. Although one difference can be found that in the modern era fuel, the fuel that is used to power machines is limited whereas the data that we have today is infinite and never ending. It is being amassed each and every second 24 X 7, 365 days and is abundant. Like every other fuel data is also useful and usable only after refinement or as we call it in technical terms 'data preprocessing'.

Like oil, data can also lose its value and be harmful in terms of a bad and misleading data analysis result if spillage happens, that is, when right data is processed and analyzed at the wrong place or when correct and noise less data is used to analyze a situation that is

wrong and absurd. Therefore, it is very dangerous to avoid or ignore the security and proper handling of data, as if mishandled it can lead to militancy in data and like oil spills it can cause serious damage to our work for which that data was collected and was being processed.

Records tell that the first camera photograph or image in the history of mankind was clicked in 1826. From that era humans and science and technology have taken a giant leap and due to that, today in 21st century images or photographs are being captured using cameras and lenses that are better than ever and are super handy, portable and easy to operate. All the advancement has led to images or photographs being a vast collection and thus becomes valuable data set for various purposes.

Today, images are used in the field of medicine to perform diagnosis and treatment of various ailments in the human and animal body. Image based checks are used to monitor the condition of the outer body of vehicles ranging from big Boeing 737s to small four wheelers. It is being used in factories to compare the newly manufactured products with a tested sample to ensure quality of the factory outputs. Everyone today knows that CCTVs are playing an important role in ensuring proper law and order situation in cities and Images have made the tiresome check-ins at airports a seamless automatic process. Today more than 80% of the images are clicked using digital portable cameras that have all kinds of data like timestamp and location data attached to them for future reference. So that when that image is accessed anytime in future the timestamp and location of the image does not remain disputed.

I.II Motivation

Humans have always appreciated images when it comes to understanding, in comparison to other expression methods like a text. They have now emerged to be the main information and data carrier in this era of science and technology. It is a better and effective means of communication for humans which includes people engaged in various types of affairs ranging from a person of humble background who has very limited access to sources to a scientist. It provides more efficient processing and also it is easier for memory recall. Images are known to have an appeal that attracts a wide range of audience and at the same time it allows those users from different backgrounds to process information accumulated in the images, in a way that's natural and usable to them.

In today's world where each smartphone comes with a handy inbuilt camera attached to the phone, with no extra costs, the surge in amount of images that is around us is indefinite. Thus, it is a very abundant and widely available data for developing deep learning models. There are approximately around 3.5 Billion smartphone users across the globe with 299.24 million being from India. This has given almost everyone the ability and possibility to record, store, and share an enormous amount of digital images. This has successfully created a large and diverse data set for creating deep learning models using computer vision techniques.

Image forensics is now a special and crucial process in many investigations. It is an investigation method by which image processing experts use image data to build a model that is used to extract valuable evidence for an under investigation case. It is even used to support blind investigations. It often uses image processing and analysis tools to recover information of strategic importance about the history of an image. Thus, in order to have such fast, reliable and scalable systems or techniques to be in place for digital forensics there is a

constant need to develop machine learning based deep learning models that are trained appropriately to deliver required data hidden in the input image and help in collecting evidence during investigations. A caption generator machine of a kind presented in this report would emerge as a valuable resource for the digital forensics field as it will be able to describe a captured scenario in human language in an unbiased and unprejudiced manner which can be used to verify the statements of accused over any particular scenarios. Since the resultant caption would be in a human readable language it would not require any technical expertise on the part of the end user, which in this case would be the investigation officer.

Everyone today is aware of the search engine Google and how it's effective searching capabilities over the year has made it our teacher for all kinds of situations. Google introduced 'Google Image Search' service in which it facilitates the internet users to search for image content over the ubiquitous World Wide Web. The user is required to describe in his/her own language an image that he/she is searching for on the web. Here the algorithm that is used reverses the process of what this report describes. It first takes a caption or an image description as an input from the user and then google repository is searched for various images that match that description entered in the search engine.

After introducing the service of 'Google Image Search' Google also introduced reverse image search functionality, extending the image search capability of the search engine. This functionality takes the image and analyzes its features like color and texture etc. After that query is generated with the help of those features and which is matched with the images in Google repository and the result is returned. The solution presented in this report when scaled up to sufficient level can be used for the betterment of reverse image search

algorithm as it uses RNN approach to generate the caption and also LSTM is used in which the image search can be linked with previous search and multiple images can be searched.

Today's fast moving world that is becoming as smart as it can day by day is also seeing an increase in law and order situations across the globe and quite unsurprisingly the culprits are also using modern tools and it is becoming a challenge to catch them. Thus, it is rightly said that 'Modern Problems Require Modern Solution', as the deep learning caption generator model presented in this report could be used in integration with the CCTV cameras installed for surveillance in different parts of cities and various public places like airports and railway stations. The problem with CCTVs is that, it requires a person constantly to look over the recordings and detect any suspicious activity, so that any law and order situation can be avoided. There is a very high possibility of human error in such a scenario. Also alarm raising in such a system is difficult which leads to delay in emergency response by suitable authority.

When the image caption generator is used with CCTVs or any other surveillance device it will do analysis and generate a human readable caption of the image it sees. This machine generated description can then be used to analyze the situation in front of the camera and also raise alarm if necessary to do so. But the system needs to be very fast and reliable as for such purposes any delay or false alarm can be very risky and even cost life and property of innocent people.

Apart from various technological motivations the presented solution is also motivated by some social responsibilities and applications. The concept of image and then describing that image in a natural language seems to be useful for blind people also. This system can be integrated with some suitable hardware that is easy to carry by such differently abled people.

As and when these people will walk through roads the scene in front of them can be captured and a real time caption generation on the spot can take place. This will create a description of any hurdle or type of terrain in front, so that the person can listen to it and decide his/her next step according to the description or caption generated.

In this era of IoT, where everything needs to be 'smart', the vehicles that we use for our commute on the road have also become smart. More and more research is being done on operating driverless vehicles or

I.III Technology Stack Involved

In order to generate captions or descriptions for images from machines, it is obvious that the machine needs to be able to correctly and reliably interpret the input image and then generate a sentence in a natural language that not only makes complete sense but also precisely describes the image for which it was generated. All of these requirements indicate that we need 'Machine Learning' to come to our rescue in order to develop a solution for the problem. This analysis is fine and seems to work well when thought initially without much stress on the implementation and model training part but as one starts looking deep into the data set that will be required to train the machine to generate these outputs, it becomes clear that we need actually need a subset of the field of 'Machine Learning' to develop the model, which is known as 'Deep Learning'. Before looking into this subset of 'Machine Learning', a better understanding about the scope and viability of machine learning is required.

Machine learning is basically making our systems or machines learn to do tasks without having to provide them explicit command for execution of that task. Today machine learning is being used in various sectors to help the people, like it is being used in the

Financial Services where banks are extensively using machine learning to perform data analytics to improve their service experience and to design the policies for their customers.

Banking frauds can also be prevented by implementing machine learning algorithms. Similarly, in the HealthCare sector machine learning has provided some very efficient lifesaving tools like sensors and wearable devices that monitor a patient's health condition. In the transportation sector the use of machine learning has changed the way people travel across the world. The transporters can now decide the most viable and demanding travel route and provide services accordingly to the commuters. The Governments across the world are using machine learning techniques to draft public health policies and perform impact analysis of their citizens and also to identify theft and analyze the aspects of national security.

The model presented in this report uses a deep learning approach because the solution required in this case although needed the use of algorithms that function similarly to machine learning but at the same time it required the use of those functionalities to be implemented across different levels. For example, the model required the use of LSTM and RNNs to select the most suitable words to describe an image from its vocabulary. Each level requires a different interpretation of the data and in fact should be an enhancement to the previous iteration. The data set that is used to train the models are somewhat different in the case of machine learning models and that of its subset, deep learning model because more structured data is needed in case of training machine learning models whereas in deep learning a deep learning the emphasis is on the Artificial Neural Networks layer. Which in this case is Recurrent Neural Network. The need of recurrent neural networks can be understood by trying to solve a problem of developing a model in which the machine is able to predict what is going to come next in a video that is being played on a device. This problem cannot be

solved by using the conventional neural networks so the need of recurrent neural networks comes into play. The RNNs can also be employed in generating deep learning models like music composition, speech processing, calligraphy designing etc. The RNNs can successfully process the variable or non-constant sequence of inputs provided to them, enabling it to be used in practical real time applications like grammar checking, image caption generator etc.

The deep learning model presented in this report uses Long-Short Term Memory (LSTM) network which is RNN with the same architecture although the function used to compute the hidden value is different from the one used in RNNs. They have cells that have the responsibility to regulate, what to keep and what to discard from the memory with each step forward. This feature makes LSTMs suitable for computing dependencies that are too long, like the one presented here, i.e., generating long sentences describing an image correctly in a natural language.

They are used to solve any problem where the previous frame or version of the input is important to generate the current version of the input and in turn this becomes relevant for further outputs. It is like a task of predicting a movie scene that is going to be played based on what just happened there. The question that why do one need a LSTM instead of RNN could be explained with the help of an example, consider trying to predict the last word of the sentence "Birds were flying in the...*sky*", here the word 'sky' is predicted and it is observed that anymore context is not required and it is quite clear that the last word would be 'sky'. In such cases RNNs can be used as the context needed is not much here and the difference between the relevant information from past and the place at which it is required is not much and relatively small. It can be deduced from the above example that a fairly low amount of context is needed here, so RNN can be used.

Now, try predicting the last word in following sentence “I joined an English speaking course….now I speak fluent *English*”, here in this example it was clear that the last word is going to the name of any language but to deduce the fact that which language the person is talking about it is very important to consider the previous line also. In such cases LSTM has to be used as context needed here is more and also the difference between the relevant information and the place at which it is required is more and relatively large. Thus, it can be deduced that when the context needed to make a decision is large, LSTM has to be used in place of RNNs.

The nature of the problem discussed in the report is such that it requires the use of highly unstructured data, i.e., image data. Along with image data the training set has five texts written in English associated with it, which describes that input image. The very first and most important problem that arises is, how we store this data in the machine so that the machine can read and understand the image. The images that we see are nothing but a collection of pixels and these pixels combine in different colors and intensity to form an image. This feature of images is used to store them or represent them in machines. So that machines can interpret the images and understand them, something that is popularly known as ‘Image Feature Extraction’.

Another field of study that is very important to mention after talking about the image data and its feature extraction is called ‘Computer Vision’, the model presented in this report uses the principles and concepts of this field. Computer vision division of artificial intelligence came into picture when the electronic devices around the world started to have the desire to see, i.e. when taking photos and videos and sharing them over platforms like

Instagram and YouTube became important and widely popular practice across the globe. The internet which is mostly built from a combination of text and images.

Text processing is relatively easier with embedding and the amount of research work done in textual sentiment analysis. Although searching and processing images requires a greater effort, as the algorithms need to know what the image wants to convey or its content. This is the exact goal that is bound to be achieved through computer vision, which can be considered as a subfield of artificial intelligence and machine learning both.

However, it is important to mention that image processing and computer vision should not be thought of as being one and the same. They have basic differences, image processing is the task of playing with images and creating a new image from an existing image using desired changes and properties. It focuses on enhancing or simplifying the image content that is already available, this simplification could be for a purpose of deep analysis or for extracting valuable features from an image. On the other hand Computer Vision can use image processing as a step in the beginning of analysis to process raw data and get it in usable format for further processing. Examples of computer vision applications include building applications to enhance the dark images and make them visible, making images noise free and also the model described in this report uses computer vision so that the contents of the image can be described in natural language based on what the computer or machine is able to ‘see’ or visualize in the input image.

Thus, this report presents a model that uses computer vision, machine learning, LSTM and RNNs, all combined to process the text descriptions and the image feature vectors and then combine them to form a single tensor, all of this happens on interactive and useful tools and python language.

I.IV Problem definition and Objectives

Computer vision, Image processing and machine learning have become very crucial needs and also economical in various fields and applications. These include applications ranging from signature recognition for authorization to iris and face recognition in forensics. Also their combination is being widely used in military applications across the world. Each of these applications has its special basic requirements, which may be unique from the others. Any stakeholder of such systems or models is concerned and demands their system to be faster, more accurate than other counterparts as well as cheaper and equipped with more extensive computational powers. All the desired traits from the systems are desirable as most of these systems are being used for mission critical purposes and scope of any mistake should be very less. Such systems are required to handle the complexity of problems of the modern world like intelligent crimes, smart city needs like smart traffic control systems, disaster control and management systems etc., thus a computer vision based model that is unbiased and free of any prejudice towards anything or anyone is required to generate a caption describing the images given to it as input. So that such description can be used to automate existing systems like traffic control systems, flood control systems or surveillance systems, this will reduce chances of errors in such critical works and also the surveillance can be conducted 24X7 without human interaction.

The problem puts forward a task of extracting features from a digital image and then describing those extracted features in a natural language.Through the localization and description of salient regions of images using LSTM a meaningful sentence in natural language will be formed that will describe images. Given a set of images and prior knowledge about the content find the correct semantic label for the entire image(s).

The major objectives of the project are :

1. Using Long-Short Term Memory (LSTM) to generate sentences in natural languages that will combine words from vocabulary based on the different focus areas of the input image and make sure that the words in the sentences are all related and makes perfect sense.
2. Demonstrate successful use of Recurrent Neural Networks (RNNs) to generate captions for input images in the system in natural language (English).
3. To get high accuracy in correctly describing an input image.

LITERATURE SURVEY

Chapter II

LITERATURE SURVEY

Kelvin Xu et.al[1] have proposed a model for automatic caption generation which is developed by the combination of all recent work done in the field of machine translation with those done in the area of object detection using computer vision. They have explored various features of an image based on attention model i.e. a model in which each word of the image description is generated focusing different areas of the image and progressively picking words from vocabulary that best describes the area focused (attention area). The model described in this paper is also self-capable of learning to amend its gaze on important entities that are present in the image while generating the sentences of description in which every word is related and makes complete sense. They further describe two types of attention mechanism-

1. Hard Version- In this version attention is focused on smaller areas of the image also.
2. Soft Version- In this the attention is more focused over an area for single word generation.

Authors have tested accuracy of the developed model using benchmark datasets like BLEU and METEOR.

Tanti et.al[2] have discussed about a model for image caption generation that rely on neural models however instead of retrieving image descriptions (either partial or wholesale) they have generated new captions by using a recurrent neural network which is usually a long short term memory (LSTM). Normally, such models use image features separated from a pre-trained convolutional neural system (CNN). For example, the InceptionV3 CNN to predisposition the RNN towards examining terms from the vocabulary so that a grouping of such terms produces captions that are applicable to the picture. The state-of-the-art work puts forward two views of how the RNNs will be used, they are as follows:-

1. Non-memory based RNNs- In this view a RNN settles on which word is mostly on the way to be generated straightaway, given what has been generated previously. In multimodal generation, this view supports models where the picture is consolidated into the RNN alongside the words that were produced so as to permit the RNN to make visually informed predictions.
2. Memory based RNNs- This subsequent view is that the RNN's job is absolutely memory-based and is just there to encode the grouping of words that have been produced this far. In this sort of setting the image description at the later layer of the RNN fills in as a blend of both the perceptual features (like words linkability index) and RNN encoding. This view empowers designs where vision and language are united late, in a multimodal layer.

Bernardi et.al [3] have ordered the current methodologies of caption generation dependent on how they conceptualize this issue, namely, models that give depiction a role as either generation problem or as a retrieval issue over a visual or multimodal illustrative space. It gives a point by point audit of existing models, featuring their advantages and disadvantages. In addition, it likewise gives an overview of the benchmark image datasets and the assessment measures that have been created to evaluate the nature of machine-generated image descriptions. It likewise extrapolates future bearings in the region of automatic image description generation. The authors finish up from the review that in contrast with the conventional keyword based image annotation (utilizing object acknowledgment, attribute discovery, scene marking, and so forth.), automatic image description frameworks produce increasingly human-like clarifications of visual content, giving a progressively complete picture of the scene.

Vinyals et.al [4] have developed a model called NIC in which they showcase an end-to-end neural network model that can automatically see a photo and produce a reasonable description in English. NIC depends on a CNN that encodes a picture into a smaller portrayal, trailed by a recurrent neural system that creates a corresponding sentence. The introduced model is prepared to amplify the probability of the sentence given the picture. Analyses on a few datasets show the strength of NIC as far as subjective outcomes (the produced sentences are truly sensible) and quantitative evaluations, using either BLEU or ranking metrics, a metric utilized in machine interpretation to assess the nature of generated sentences. It demonstrated that as the size of the accessible datasets for image description increments, so will the exhibition of approaches like NIC.

Kuznetsova et.al [5] have presented an all encompassing data-driven way to deal with image description generation by exploiting the huge measure of (boisterous) associated natural language descriptions accessible on the web and parallel image data. The model retrieves a current human-made expression used to depict outwardly comparable pictures,given a query image, at that point specifically join those expressions to produce a novel description for the inquiry picture. It gives the generation procedure a role as constraint enhancement issues, altogether joining various interconnected parts of language composition for content arranging, surface acknowledgment and discourse structure. Assessment by human annotators shows that their last system produces more semantically right and phonetically engaging descriptions than two nontrivial baselines.

Siming Li et.al [6] have presented an essential yet a compelling method to manage automatically forming image descriptions given PC vision based wellsprings of information and using web-scale n-grams. Unlike most past work that layouts or recoups past content relevant to a picture, this method structures sentences totally without any planning. Test outcomes show that it is functional to deliver fundamental straightforward depictions that are applicable to the specific content of a picture, while permitting inventiveness in the description – making for more human-like explanations than previous strategies. This strategy involves two phases: (n-gram) phrase fusion and (n-gram) phrase selection.

1. Phrase selection – assembles candidate phrases that may be possibly useful for delivering the depiction of a given picture.

2. Phrase fusion – finds the perfect good arrangement of phrases using dynamic programming to make another (and progressively erratic) state that delineates the picture.

Ryan Kiros et.al [7] have proposed two multimodal neural language models: models of trademark language that can be adjusted on various modalities. An imagetext multimodal neural language model can be used to recuperate pictures given complex sentence inquiries, recoup phrase descriptions given picture queries, similarly as to produce content adapted on pictures. In contrast to a critical number of the present procedures, this technique can create sentence descriptions for pictures without the usage of configurations, organized prediction, and moreover syntactic trees. Or maybe, it relies upon word portrayals picked up from countless words and molding the model on high level image features picked up from deep neural systems. They introduced two systems subject to the log-bilinear model of Mnih and Hinton (2007): the factored 3-way log-bilinear model and the modality-biased log-bilinear model. Word portrayals and image features are discovered together by commonly setting up our language models with a convolutional network.

Yang et.al [8] have suggested a term creation technique that explains images by anticipating the most possible verbs, propositions, nouns and verbs that make up the central sentence structure. The input is the original noisy measurement of scenes and objects found in the picture using trained detectors which are state of the art. As the direct estimation of behavior from still photographs is inaccurate, a language model is used to train the English Gigaword corpus in order to achieve their estimates; along with the probability of co-located nouns, scenes and prepositions. These projections are used as variables for the HMM that represent

the phrase generation process, with secret nodes as phrase components and image detection as pollutants. Description of an image is the result of an incredibly complex process that involves: 1) interpretation in Visual Space, 2) foundation in World Experience in Language Space, and 3) speech / text creation. Experimental findings indicate that this technique of integrating vision and language creates coherent and concise sentences relative to simplistic approaches that use vision alone.

Zaremba et.al [9] have provided a brief regularization technique using Long Short-Term Memory (LSTM) from Recurrent Neural Networks (RNNs). Unluckily, for RNNs the most powerful regularization method for feed-forward neural networks does not work. As a consequence, realistic implementations of RNNs sometimes use versions that are too tiny whereas big RNNs appear to be over-fitting. Present regularization approaches have fairly minor changes for RNNs Graves (2013). Through this work, the authors show that dropout, when it is used correctly, the overfitting in LSTMs significantly reduces and assesses it on three different problems.

Barnard et.al [10] have provided a different method for simulating multimodal data sets, based on the particular case of segmented photographs with corresponding text. There are many programs to learn about the joint distribution of visual features and words.

They find in depth the identification of terms identified with entire images (self-annotation) and referring to specific picture regions (region naming). Auto-annotation can organize and view a huge collection of photographs. Region labelling is a model of object recognition as a process of interpreting regions of images into words, which can be translated from one

language to another. Learning the association between image regions and semantic correlations (words) is an excellent case of multimodal data mining, especially as it is usually difficult to apply data mining methods to picture collections. Notably because data mining methods are generally difficult to apply to the series of images. They establish a variety of models for the mutual distribution of picture regions and words, including those that directly study communication between regions and words. The authors analyze multi-modal and relationship modifications to Hofmann 's hierarchical clustering / appearance scheme, a translation mechanism that developed from a mathematical computer translation (Brown et al.), and a multi-modal extension to a mixture of latent Dirichlet allocation (MoM-LDA). The models are tested utilizing a wide set of annotated photographs of actual scenes.

Yao et.al [11] have presented an Image to Text (I2 T) framework that converts image and video content to text descriptions based on image (or frame) comprehension. The suggested framework is accompanied by three steps. The input images with the use of an image parsing engine are fragmented in their visual features in the first stage. In the second stage, the effects of the first stage are transformed into a textual description in the context of the Network Ontology Language (OWL). Finally, in the third stage, the OWL representation of the preceding step is transformed into semantically meaningful, human readable and searchable text reports by a text generation engine. Visual information representation from And-or-Graph (AoG) is the key component of the I2T system. It offers a visual representation of a large-scale image for the learning of a categorical photo and symbolic representation. During image parsing it takes a top-down approach and binds low-level image features with high-level semantical conceptions so that the picture can be translated into semantical

metadata and eventually into a written text. The I2 T framework is particularly different in that it generates annotations that are semantically meaningful. Since the content of the image and video is converted into both OWL and text format, this framework can be merged with a full text search engine to provide error-free content-based recovery. Users can also query images and video clips based on keywords and semance.

Kumar et.al [12] have used deep learning to propose an Image Caption Generator. It aims at producing captions based on processes that require both image processing and computer vision for an input image. The method identifies the connections in the image between various people , objects and animals whilst capturing and turning semantic meaning into a human natural language. Regional Object Detector (RODe) is used to identify, recognize and produce captions. The method proposed is centered on deep learning to further improve the existing system. This method is applicable to a dataset of 8k Flickr. It created captions with a more concise meaning and detailed importance than the current generators of image captions.

Shabir et.al [13] have deduced that, since the research community has been very interested in finding new ways to automate content-based image retrieval, they have presented an overview of some technical aspects and techniques for caption-generation of images. The paper discusses briefly the description of certain new research and also the relevant points of ongoing work. There are various picture description frameworks but findings indicate that a better architecture with improved results is still needed. Four main fields are identified, where potential efforts to enhance results should be carried out. Firstly, methods used to create several sentences will consider various background environments. Of improved output, the

efficiency can be increased by a mixture of different models of association with the caption. Secondly, there must be different measurement criteria for improved output of the summary generation method, but do not affect the framework due to lack of extensive annotation. Thirdly, a framework needs to exist that creates identical multiple descriptions of a picture with specific contents. Fourthly, estimation of tiny items is always a concern owing to poor judgment. Additionally, annotation expressions created by high-level definition will improve efficiency.

Li et.al [14] have incorporated the visual storytelling issue into the literature. It aims to produce coherent and concise lengthy-video stories. Video storytelling poses new challenges because of the variety of the plot, the length as well as the quality of the images. The writers suggest innovative approaches for overcoming the problems. To begin with, the authors are proposing a context-aware paradigm for multimodal embedding research, they are designing a Residual Bidirectional RNN to use past and future contextual knowledge. Multimodal embedding is then used to access video clip phrases. Secondly, they suggest a Narrator model for choosing excerpts that explain the specific plot. The Narrator is intended as a reinforcement learning agent who is trained by explicitly improving the generated story's textual metric. On the Video Story dataset, they access the technique, a data collection that they have collected to allow analysis. They equate the approach with different state-of-the-art baselines and demonstrate that their methodology works well in terms of objective measurements and usage analysis.

Xiangyang et.al [15] have presented a system for image captioning based on scene graphs has been suggested by the writers in this paper. A few techniques have recently captured based on semantic ideas from photographs and converted them into high-level depictions afterwards. While considerable improvement has been made, the majority of prior approaches individually handle individuals in pictures, thereby missing organized information that offers meaningful signals for image captioning. Scene graphs provide a significant volume of ordered details, since they not only represent objects in pictures but often include pair-by-pair connections. CNN features from the bounding box offsets of entity instances are derived for visual representation, as well as semantic interaction features from triplets (e.g. man eating fruit) for semantic representation, to include both image features and conceptual knowledge in structured scenario graphs. After acquiring these characteristics, the authors present at each step a hierarchical attention-based module for learning discriminative highlights for word generation.

REQUIREMENTS ANALYSIS

Chapter III

REQUIREMENTS ANALYSIS

III.I Functional Requirements

Functional Requirement No.	**Functional Requirement Description**
FR 1	The dataset used should be from a reliable source and in proper format.
FR 2	The input image should be used to generate a 14 X 14 feature map using convolutional feature extraction. This should reduce the dimensionality

	of input images so that they can be used in RNN with attention over the image.
FR3	Use two different attention mechanisms over the input image, namely stochastic attention and deterministic attention and combine the result to generate words for the caption.
FR 4	The splitting of available dataset into test data and training data should be proper and a balanced split to produce a good classification model.
FR 5	Long short term memory (LSTM) should be incorporated along with RNNs so that one word can be generated describing the current focus area but it should be conditioned on the previously saved state in the memory.
FR 6	Examine and interpret the image focus areas against the "where" and "what" conditions of the attention mechanism.
FR 7	Perform validation of usefulness of attention quantitatively in the task of generating captions with respect to performance on Flickr8k dataset.
FR 8	The test of the trained model on holdout set should be performed and result in maximum accuracy.
FR 9	The model should be able to generate captions for new photographs that were not present in the data used to train it.

III.II Non Functional Requirements

Non-Functional Requirements No.	**Non-Functional Requirements Description**

NFR 1	Response Time of the developed model should be less and convenient for users.
NFR 2	The model should be Scalable, that is, it should work for large data also with the same deliverables.
NFR 3	The model should have high usability and interoperability.

III.III Software Requirements

Python-3.7.4:

It is a programming language that is interpreted, high level and of general purpose. This language is chosen for this project because of its high data handling capacity and also its simplicity.

Python pickle:

The Python pickle module is used to serialize (process of converting an object into byte stream) and deserialize (process of converting a byte stream into an object) the Python object structure. Every object in Python could be pickled so that it can be stored on the disk. This is required because it provides some degree of persistence in data since it is stored on disk, so

that it can be used later. We have used this module to easily save and reload the deep learning models.

Nltk.translate.bleu_score package:

Returns an associated sentence entity that encapsulates two sentences together with an alignment between them. In machine translation, this is usually used to represent a sentence and its translation. The Bilingual Evaluation Understudy Score (BLEU) is a benchmark for assessing the sentence produced by the comparison sentence. The method works by measuring corresponding n-grams in the nominee translation to n-grams in the comparative language, where 1-gram or unigram will be each symbol, and a bigram relation will be each word pair and a contrast would be rendered irrespective of the word order.

NumPy-v1.17:

NumPy is a key package for scientific programming in Python. It is a library for Python that is used for scientific computing and mathematical data. It offers things such as large multidimensional arrays and matrices. It also provides a variety of routines for fast array operations, including statistical, logical, and form manipulation.

Matplotlib-3.1.1:

It's a Python programming language and NumPy plotting library. It integrates plots into application through an object-oriented API. It can plot 2D graphs for arrays. It is extremely handy because it helps in visualization of the data which helps us gain useful insights.

Keras:

It is an open source Python library which allows us to build deep learning or machine learning algorithms. It is easy to use since it provides an abstract interface to build the models. It is compatible with versions 2.7-3.7 of Python.

III.IV Hardware Requirements

Operating System: Windows 10

Processor: 1.5GHz processor

RAM: 8 GB

SYSTEM DESIGN

Chapter IV

SYSTEM DESIGN

A typical recurrent neural network encoder-decoder design is used to solve the question of creating image captions. It has two aspects involved in it:

1. **Encoder**: A network architecture that uses an internal representation to interpret the picture data and translate the information into a fixed-length vector.
2. **Decoder**: A network model that reads the encoded picture and produces the output as a textual interpretation.

The merge model is a type of encoder-decoder architecture as defined in Figure 1. The merge approach integrates both the encoded image input form and the encoded text summary form that has been produced so far. A very basic decoder model then utilizes a combination of

those two encoded inputs to produce the next term in the chain. The method only uses the RNN to encode the text created up to now. This distinguishes the issue of processing the input of pictures, the input of text and the mixture and perception of the encoded data.

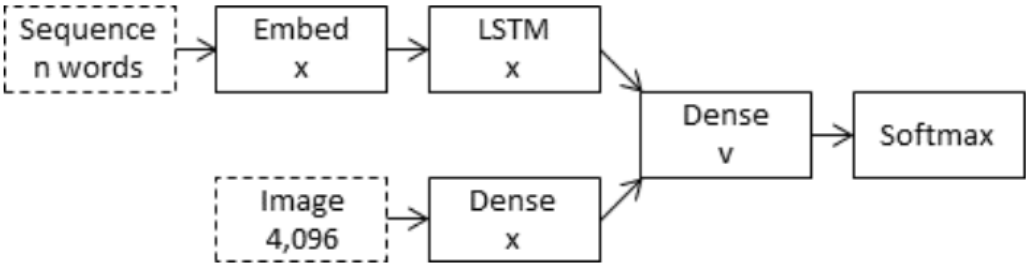


Figure 1: Merge Model

The proposed system is based on the "merge-model". The schematic of the model is reproduced in Figure 2.

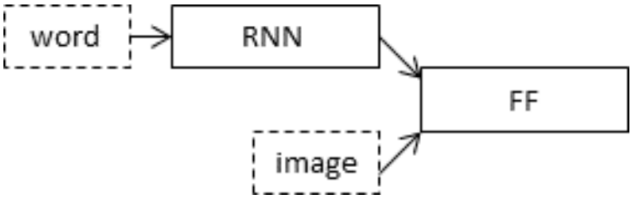


Figure 2: Merge Model Schematic

The deep learning model development can be represented in three parts:

- **Photo Feature Extractor:** This is an InceptionV3 layer that is pre-trained on the ImageNet dataset. The images will be pre-processed using the InceptionV3 model

without the output layer and will use the derived features projected by this model as data.

- **Sequence Processor:** It is a word embedding layer for managing text data, preceded by a Long Short-Term Memory (LSTM) recurrent neural network layer.

- **Decoder:** Both the sequence processor and feature extractor produce a fixed-length vector. They are bundled together and analyzed by a Dense Layer to reach a final prediction.

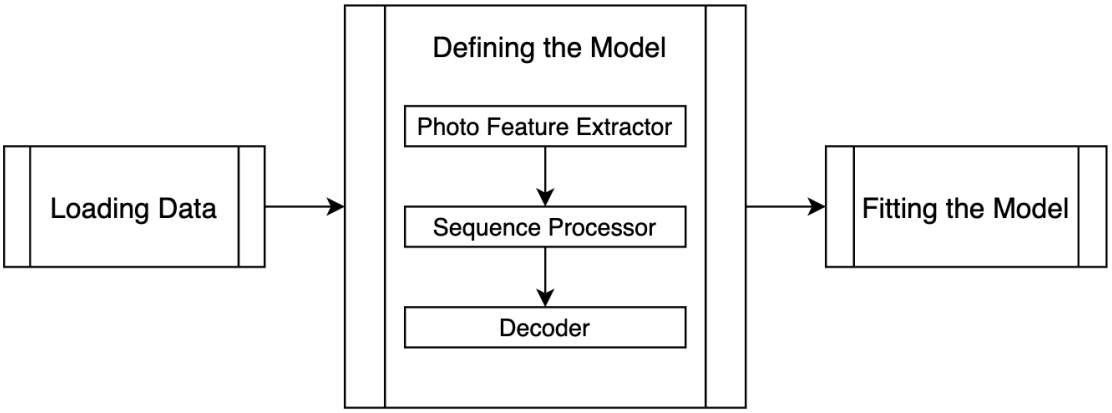


Figure 3: Development of Deep Learning Model

The Photo Feature Extractor outputs the input image attributes into a 4,906 matrix. A 256 element visual representation is then achieved when the attributes from the Photo Feature Extractor are fed into a Dense layer. The Sequence Processor model needs the input variables

to be of a preset value, in our case, that will be 34 words which is the length of the biggest caption in the dataset. These variables are inserted into the Embedding layer and then later into an LSTM layer of 256 processing units. Both the models will output a 256 element matrix. Both the models use a 50 percent dropout regularization to limit the amount of over fitting since this model learns very swiftly. Finally, the Decoder part of the model combines the elements from both the input models by using an addition operation. The output from the addition layer is fed into a 256 element Dense layer and then finally into the final Dense layer which outputs the softmax project about the whole sample vocabulary to predict the next term in the sequence of the caption description.

Figure 4 provides a plot to represent the network structure that helps to better explain the two input flows.

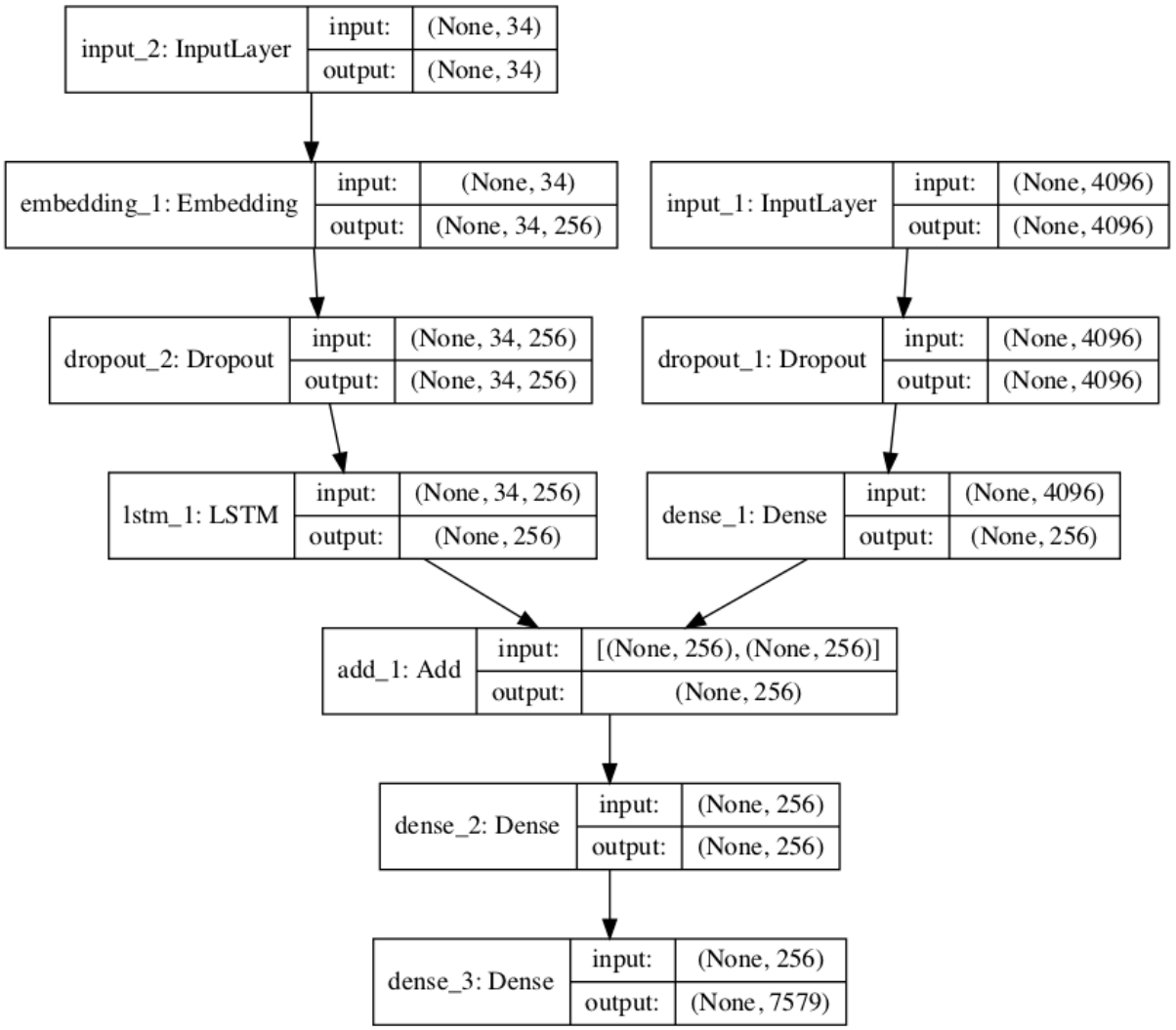


Figure 4: Network Structure

Figure 4 demonstrates the flow of two input streams in the model as it moves through the different LSTM units of the Recurrent Neural Network (RNNs) to eventually evaluate each word in the vocabulary that defines the current target area for the input image, thus taking into consideration all the terms previously used when producing the caption in the natural language.

The overall data flow diagram of the proposed system is shown in Figure 5.

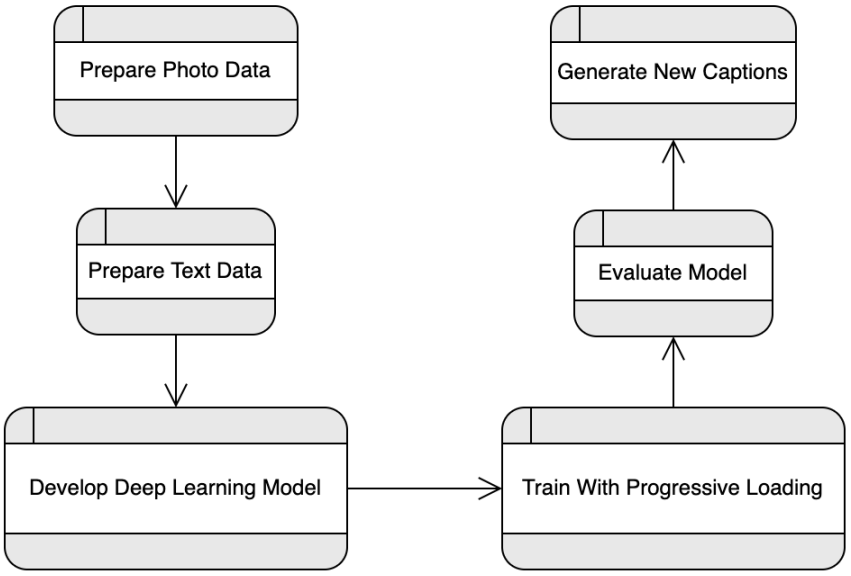


Figure 5: Data Flow Diagram

The data flow diagram depicts the flow of data across our proposed model which starts with preparing the photo data collected from different sources and preprocessing them to be in expected input format. After this the text descriptions (captions) associated with each of the images are prepared and both the text and image data are combined to develop the deep learning model using a combination of photo feature extractor, sequence processor and decoder, which is followed by training the developed model using progressive loading to ensure accuracy. Evaluation or test of the developed model is performed using the test data splitted from the total training set. Once the satisfactory or desired accuracy is reached the model can be used to generate captions for the new input images.

IMPLEMENTATION

Chapter V

IMPLEMENTATION

Generating a textual depiction for a given photo is a challenging artificial intelligence issue. It requires two strategies; one is to understand the content of the picture and another is a language model (from the field of NLP) which transforms the content of the picture into words organized appropriately. These deep learning strategies have exhibited state-of-the-art results on caption generation issues. These techniques have a solitary end-to-end model which can be characterized to foresee a textual description, given a photograph, rather than requiring advanced data preparing or pipeline of explicitly planned models.

V.I Data Collection

With the end goal of this task, MS-COCO dataset is utilized. The dataset contains 82000 pictures and each image is associated with at least 5 captions. To quicken the training speed, a subset of 30000 captions and the associated images are randomly selected to train the model. It is sensible and moderately little so it very well may be utilized to run on workstations utilizing a CPU as opposed to depending on a GPU of a powerful machine. Thus, it is conceivable to run a system that is certainly not a very good quality PC/Laptop. The pictures are split as follows:

1. Training Set (24000 pictures)
2. Test Set (6000 pictures)

Each picture is combined with five distinct captions which give away from the notable entities and events. These pictures were taken from six events, and they do not contain any notable individuals or areas, however were physically chosen to portray an assortment of scenes and circumstances.

a little girl in a pink dress going into a wooden cabin .
a little girl climbing the stairs to her playhouse .
a little girl climbing into a wooden playhouse .
a girl going into a wooden building .
a child in a pink dress is climbing up a set of stairs in an entry way .

two dogs on pavement moving toward each other .
two dogs of different breeds looking at each other on the road .
a black dog and a white dog with brown spots are staring at each other in the street .
a black dog and a tri-colored dog playing with each other on the road .
a black dog and a spotted dog are fighting

young girl with pigtails painting outside in the grass .
there is a girl with pigtails sitting in front of a rainbow painting .
a small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it .
a little girl is sitting in front of a large painted rainbow .
a little girl covered in paint sits in front of a painted rainbow with her hands in a bowl .

man laying on bench holding leash of dog sitting on ground
a shirtless man lies on a park bench with his dog .
a man sleeping on a bench outside with a white and black dog sitting next to him .
a man lays on the bench to which a white dog is also tied .
a man lays on a bench while his dog sits by him .

the man with pierced ears is wearing glasses and an orange hat .
a man with glasses is wearing a beer can crocheted hat .
a man with gauges and glasses is wearing a blitz hat .
a man wears an orange hat and glasses .
a man in an orange hat starring at something .

Figure 6: Five Captions for Each Image

V.II Prepare Photo Data

A pre-trained model is utilized to decipher the contents of the photographs. We utilize the InceptionV3 model which Keras gives legitimately. The issue is, it is an enormous model

and running every photograph through the system each time another language model configuration is to be tried is repetitive. Rather, by utilizing the pre-trained model the features are pre-figured and saved to a document. These features are then loaded into the model to translate a given photograph in the dataset. It is indistinguishable from running the photograph through the full InceptionV3 model (it simply happens once in advance). This improvement will consume less memory and will make preparing the models quicker. Using the InceptionV3 class, the InceptionV3 model is stacked in Keras. The last layer from the model is expelled, as this is the model which is used to anticipate a classification for a photograph. Just the internal representation of the photograph is of significance. These are the features that the model has received from the photograph. Keras also provides the devices to reshape the stacked photograph into a desired size for the model. Given a registry name, the function named load_image() that will stack every photograph, set it up for Inception, and gather the predicted features from the Inception model. The pictures are 1-dimensional 4096 element vectors. The function returns a word reference of picture identifiers to picture features. The function is called to set up the photograph information for testing the model.

```
[ ]  # Resize to 299x299 pixels and preprocess
     def load_image(image_path):
         img = tf.io.read_file(image_path)
         img = tf.image.decode_jpeg(img, channels=3)
         img = tf.image.resize(img, (299, 299))
         img = tf.keras.applications.inception_v3.preprocess_input(img)
         return img, image_path
```

```
[ ]  image_model = tf.keras.applications.InceptionV3(include_top=False,
                                                      weights='imagenet')
     new_input = image_model.input
     hidden_layer = image_model.layers[-1].output

     image_features_extract_model = tf.keras.Model(new_input, hidden_layer)
```

V.III Prepare Text Data

The MSCOCO dataset contains various portrayals for every photo and the content of the depictions requires some cleaning. Every photograph contains a novel identifier. This identifier is utilized on the photograph filename and in the content document of depictions. The depiction of the text needs cleaning. The depictions are now simple to work with because they are tokenized. The text is cleaned in the following ways so as to decrease the length of the jargon of words the model needs to work with:

- Converting all the words to lowercase.
- Removing all punctuation.
- Removing all words that are one character or less in length (e.g. ‘a’).
- Removing all words which contain numbers in them.

Beneath characterizes the clean_descriptions() function that, given the word reference of picture identifiers to depictions, ventures through every portrayal and cleans the test.

Once cleaned, the size of the jargon can be summed up. Preferably, a jargon ought to be both expressive and as little as could be expected under the circumstances. A smaller jargon will bring about a smaller model that will prepare quicker. At last, the word tokenized vector of picture identifiers and depictions are saved into a variable named train_seqs.

```
[ ]  # Find the maximum length of any caption in our dataset
     def calc_max_length(tensor):
         return max(len(t) for t in tensor)
```

```
[ ]  # Choose the top 5000 words from the vocabulary
     top_k = 5000
     tokenizer = tf.keras.preprocessing.text.Tokenizer(num_words=top_k,
                                                       oov_token="<unk>",
                                                       filters='!"#$%&()*+.,-/:;=?@[\]^_`{
     tokenizer.fit_on_texts(train_captions)
     train_seqs = tokenizer.texts_to_sequences(train_captions)
```

```
[ ]  tokenizer.word_index['<pad>'] = 0
     tokenizer.index_word[0] = '<pad>'
```

```
[ ]  # Create the tokenized vectors
     train_seqs = tokenizer.texts_to_sequences(train_captions)
```

```
[ ]  # Pad each vector to the max_length of the captions
     # If you do not provide a max_length value, pad_sequences calculates it automatically
     cap_vector = tf.keras.preprocessing.sequence.pad_sequences(train_seqs, padding='post'
```

```
[ ]  # Calculates the max_length, which is used to store the attention weights
     max_length = calc_max_length(train_seqs)
```

V.IV Develop Deep Learning Model

A. Loading Data

The text information and prepared photograph must be stacked with the goal that it tends to be utilized to fit the model. Based on all the photographs and inscriptions in the preparation dataset, the model will be trained. The exhibition of the model will be observed on the development dataset during training and will utilize that performance to figure out when to save models to the record. The test and train dataset sizes have been defined as 24000 and 6000 respectively.

The model produces a caption given a photograph, and the caption will be produced each word in turn. The grouping of recently produced words will be given as input. Along these lines, there is a requirement for a 'first word' to start the generation procedure and a 'last

word' to show the finish of the inscription. The string '<start>' and '<end>' are utilized for this. These strings are added to the stacked depictions while they are stacked. It is critical to do this so that the tokens are encoded before the text is encoded.

Before the description text gets introduced to the model as an input or contrasted with the model's forecasts, it should be encoded to numbers. The initial phase is to map the words to unique integers. The Tokenizer class in Keras library, provides a way to load these mapping as input, from the loaded description data.

The text is currently to be encoded. Every depiction is split into words where the model is given a single word and the photograph which will produce the next word. Further the first 2 words of the caption is given as input along with the picture to produce the next word. This is the means by which the model will be prepared. For instance, the sequence "little boy walking on road" is part into 6 input-output sets to prepare the model:

```
1 X1,     X2 (text sequence),                    y (word)
2 photo   startseq,                              little
3 photo   startseq, little,                      boy
4 photo   startseq, little, boy,                 walking
5 photo   startseq, little, boy, walking,        on
6 photo   startseq, little, boy, walking, on,    road
7 photo   startseq, little, boy, walking, on, road,   endseq
```

Figure 7: Data points corresponding to one image and its caption

Later the produced words will be linked and recursively given as input to produce a caption for a picture when the model is utilized again to produce depictions. Given the tokenizer, the word reference of all depictions and photographs, a most extreme sequence length as input, the code will change the information into input-output sets of information for preparing the model. There are 2 input arrays to the model: one for the encoded content and

one for photograph features. One output from the model is the next word in the sequence which is encoded. The input text will be fed to a word embedding layer which is encoded as integers. The photo features are loaded directly to another part of the model. The model yields a prediction, which is a probability distribution over all the words in the vocabulary. The information yielded will subsequently be a one-hot encoded variant of each word, representing an idealized probability distribution with 0 values at all word positions aside from the actual word position, which has a value of 1.

B. Defining the Model

The model is characterized dependent on the "merge-model" portrayed by Marc Tanti, et al. A standard encoder-decoder RNN architecture is utilized to address the image generation issue. This included two components:

1. Encoder: It encodes the content into a fixed length vector by a network model that reads the photograph input using an internal representation.
2. Decoder: A network model that generates the textual description by reading the encoded photograph.

The merge model joins the encoded type of the text depiction produced up until now with the picture input which is also encoded. A basic decoder model is used to load these 2 encoded inputs, to produce the following word in the sequence. The methodology utilizes the RNN just to encode the text produced up until this point. This isolates the worry of demonstrating the picture input, the text input and the combining and translation of the encoded inputs.

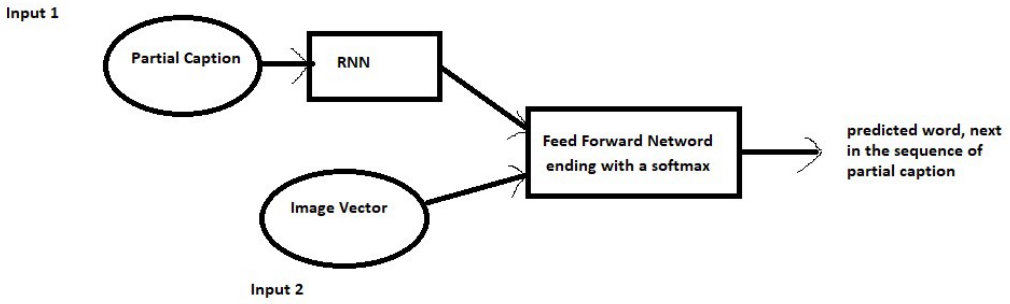


Figure 8: Schematic of the Merge Model for Image Captioning

Since the input comprises two sections, a picture vector and a fractional caption, the Sequential API given by the Keras library can't be utilized. Because of this explanation, the Functional API is utilized which permits us to make Merge Models. The beneath plot assists with envisioning the structure of the network and better comprehend the two streams of input.

The Photo Feature Extractor model utilizes photograph features as input and it expects it to be a vector of 4096 components. They are then fed to the Dense layer to deliver a 256 component representation of the photograph. The Sequence processor model requires the input sequences to be of predefined length (34 words). This is taken by the Embedding layer which utilizes a mask to ignore padded values. This is trailed by a LSTM layer with 256 memory units. The LSTM layer is only a particular RNN to process the sequence input i.e partial captions. Both the input models produce a 256 component vector and use regularization as half dropout. As the model learns quickly, this is done to diminish overfitting the preparation dataset. Further, an addition operation is performed on both the input models by the decoder model which merges them into a single vector. This is then sent

to a dense 256 neuron layer and afterward to a final output Dense layer that makes a softmax prediction over the whole output vocabulary for the following word in the sequence.

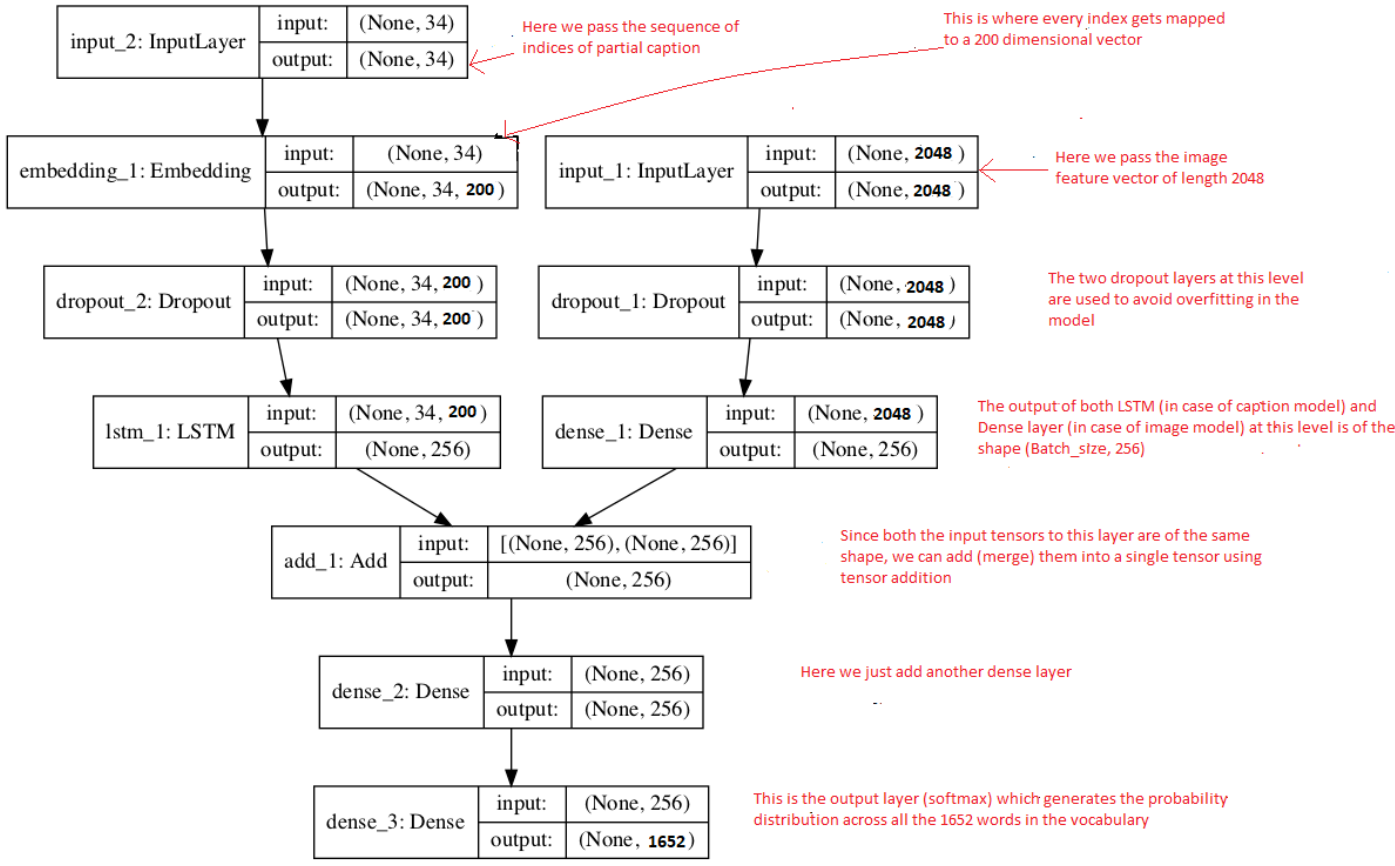


Figure 9: Plot of the Caption Generation Deep Learning Model

```
class CNN_Encoder(tf.keras.Model):
    # Since you have already extracted the features and dumped it using pickle
    # This encoder passes those features through a Fully connected layer
    def __init__(self, embedding_dim):
        super(CNN_Encoder, self).__init__()
        # shape after fc == (batch_size, 64, embedding_dim)
        self.fc = tf.keras.layers.Dense(embedding_dim)

    def call(self, x):
        x = self.fc(x)
        x = tf.nn.relu(x)
        return x
```

```
class RNN_Decoder(tf.keras.Model):
  def __init__(self, embedding_dim, units, vocab_size):
    super(RNN_Decoder, self).__init__()
    self.units = units

    self.embedding = tf.keras.layers.Embedding(vocab_size, embedding_dim)
    self.gru = tf.keras.layers.GRU(self.units,
                                   return_sequences=True,
                                   return_state=True,
                                   recurrent_initializer='glorot_uniform')
    self.fc1 = tf.keras.layers.Dense(self.units)
    self.fc2 = tf.keras.layers.Dense(vocab_size)

    self.attention = BahdanauAttention(self.units)
```

C. Fitting the Model

The model catches on quickly and rapidly overfits the preparation dataset. Hence, the prepared model ought to be observed on the holdout development dataset. At the point when the skill of the model improves on the development dataset toward the end of an epoch, the entire model will be saved to a document. Toward the finish of the run, the final model will be one of the saved models which give the best accuracy on the preparation dataset. This is finished by characterizing a ModelCheckpoint in Keras and determining it to screen the minimum loss on the validation dataset and save the model to a record that has both validation and preparing loss in the filename. The model is fit for 20 epochs.

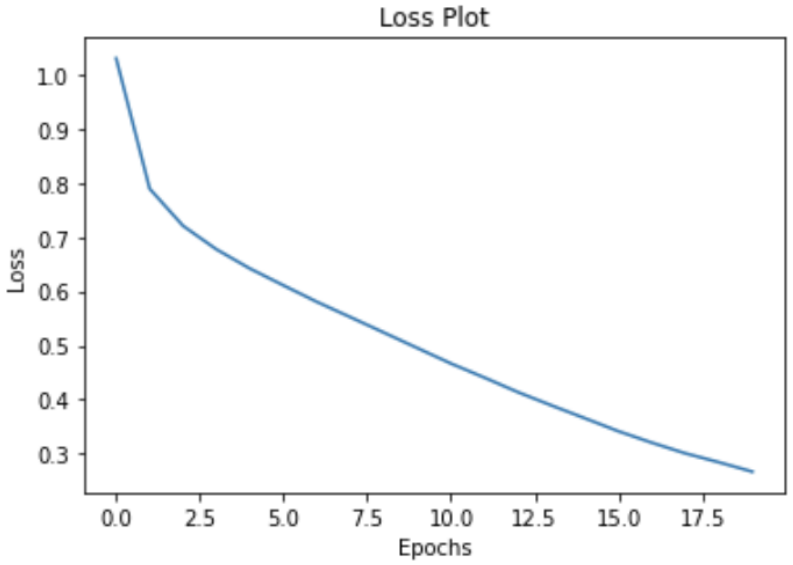


Fig 10: Loss Plot

V.V Train With Progressive Loading

This is an optional step. The preparation of the caption model devours a great deal of RAM. It isn't memory efficient. The model can't be prepared on an ordinary workstation with 8GB RAM. A workaround is to utilize progressive loading. A generator is a function that returns batches of tests for the model to prepare on.

An information generator will yield one photograph of information for every batch. This will be the entirety of the sequences produced for a photograph and its set of depictions. The absence of the unrolled sequences of train and test information in memory before fitting the model, is a huge memory saving feature it offers, yet that these examples are made varying per photograph.

TESTING

Chapter VI

TESTING

Once the model is fit, it can be assessed using the holdout test dataset. The dataset contains completely 30000 photos for our use. The division of images from our dataset for different purposes while developing the model are as follows -

1. Training - 24000 data items (i.e. images & their descriptions/captions)
2. Testing - 6000 data items (i.e. images & their descriptions/captions)

The developed model is evaluated by testing the machine generated captions for the photos in our test dataset against a standard cost function. Tests are needed to decide whether or not the model is overfitting.

Since testing the developed model involves judging the accuracy of image captions generated by the machines that are in natural language, a score or benchmark was needed to determine the accuracy of the captions. As only then the developed model can be tested whether it is able to generate the captions based on the training and secondly, whether those captions accurately describe images and meaningful sentences are generated by identifying various aspects of images. BLEU score was used as a benchmark to test the generated captions for images given as input from the test set. Details about implementation of BLEU and the results are discussed in the next chapter in detail.

The testing phase for the presented model can be summarised as follows, in most cases the model has been able to produce output in the form of captions that convey the scenario depicted in an image. However, in some of the situations it got confused to describe an event due to lack of a token in the dictionary. It also got confused in cases where one colour scheme covered almost the entire image. The final aspects of the developed model including test results are described in the next chapter.

RESULTS

Chapter VII

RESULTS

BLEU score defines the exactness of the mode. Bilingual evaluation understudy (BLEU) is an algorithm which evaluates the quality of the text that a machine has interpreted. It was one of the first to reach high correlation with human judgment. BLEU score is always defined as 0 to 1, where 0 means that the generated image description or caption is not able to describe the image features correctly at all.

1. A numerical closeness metric of the translation which is then generated for captions generated for each of the images and measured against
2. A corpus of translations by human reference.

To measure the BLEU score, the captions are generated first for all the test images, and then the captions produced by these machines are used as nominee sentences. The candidate sentences are contrasted with the human-given captions and the candidate's average BLEU score relating to each of the references. So we measure 6000 BLEU scores for 6000 test photos using the Natural Language Toolkit (NLTK) which is a python package. We compare each description generated against all photo reference descriptions. We then calculate cumulative n-grams of BLEU scores for 1 , 2, 3 and 4.

An average of the BLEU score over 6000 test images are taken. The model's net BLEU score after 20 epoch training was calculated to be **0.37** for uni-grams, **0.18** for bi-gram, **0.11** for tri-gram and **0.04** for 4-grams. The state of the art outcomes can be obtained by raising the amount of epochs and increasing the number of training data points but that will entail higher computation.

```
from nltk.translate.bleu_score import corpus_bleu
def evaluate_model():
  actual, predicted = list(), list()

  len_val = len(img_name_val)
  for rid in range(0, len_val):
    image = img_name_val[rid]
    real_caption = ' '.join([tokenizer.index_word[i] for i in cap_val[rid] if i not in [0]])
    result, attention_plot = evaluate(image)

    actual.append([real_caption.split()])
    predicted.append(result)



  print('BLEU-1: %f' % corpus_bleu(actual, predicted, weights=(1.0, 0, 0, 0)))
  print('BLEU-2: %f' % corpus_bleu(actual, predicted, weights=(0.5, 0.5, 0, 0)))
  print('BLEU-3: %f' % corpus_bleu(actual, predicted, weights=(0.3, 0.3, 0.3, 0)))
  print('BLEU-4: %f' % corpus_bleu(actual, predicted, weights=(0.25, 0.25, 0.25, 0.25)))
```

Fig 11: Code snippet for BLEU score weights for different n-grams

```
from nltk.translate.bleu_score import corpus_bleu
def evaluate_model():
  actual, predicted = list(), list()

  len_val = len(img_name_val)
  for rid in range(0, len_val):
    image = img_name_val[rid]
    real_caption = ' '.join([tokenizer.index_word[i] for i in cap_val[rid] if i not in [0]])
    result, attention_plot = evaluate(image)

    actual.append([real_caption.split()])
    predicted.append(result)



  print('BLEU-1: %f' % corpus_bleu(actual, predicted, weights=(1.0, 0, 0, 0)))
  print('BLEU-2: %f' % corpus_bleu(actual, predicted, weights=(0.5, 0.5, 0, 0)))
  print('BLEU-3: %f' % corpus_bleu(actual, predicted, weights=(0.3, 0.3, 0.3, 0)))
  print('BLEU-4: %f' % corpus_bleu(actual, predicted, weights=(0.25, 0.25, 0.25, 0.25)))

evaluate_model()
```

```
BLEU-1: 0.372867
BLEU-2: 0.186140
BLEU-3: 0.113631
BLEU-4: 0.045331
```

Fig 12: BLEU scores of the model

The developed model is able to successfully predict the captions for new images as shown below. The trained model can be loaded and used to describe the features of new images. This

implementation can be implemented as an API for external use in various applications such as websites or apps. Few examples of captions generated for new images are shown below.

Fig 13: Example 1

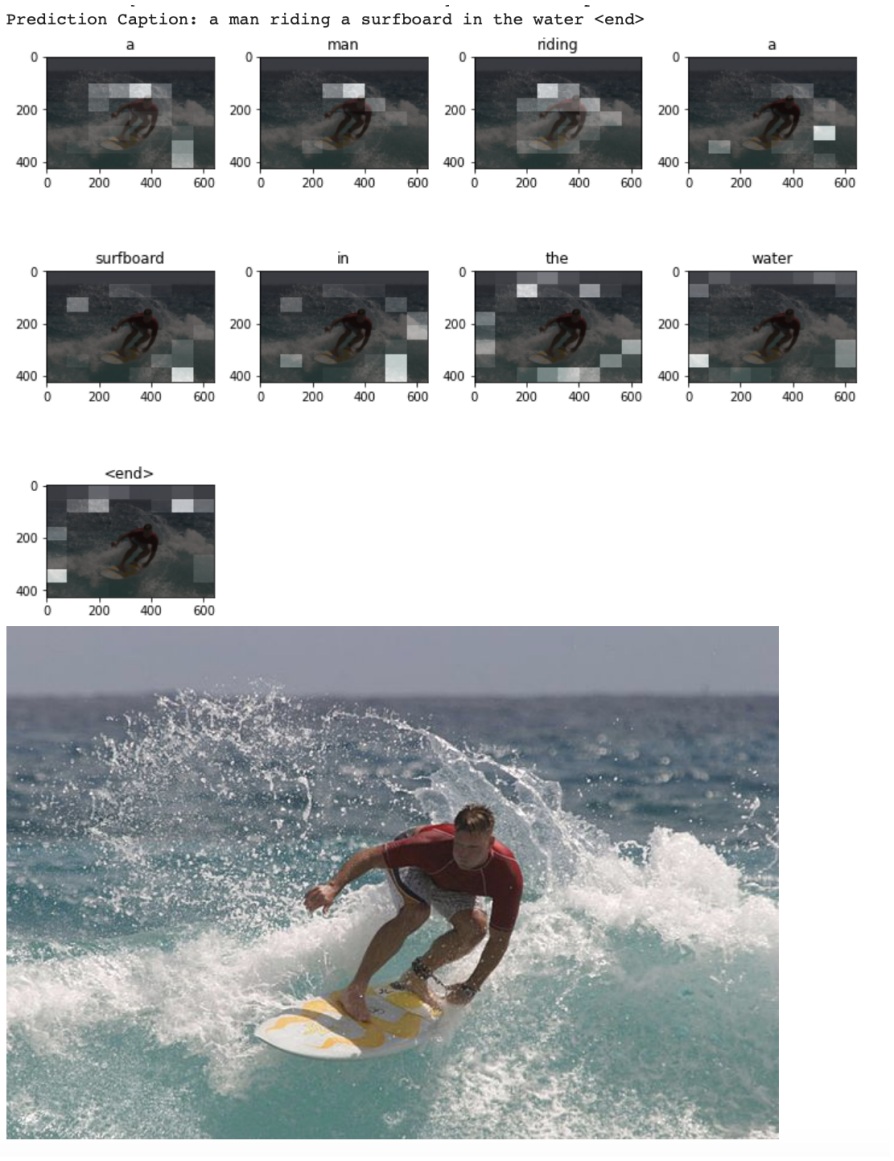


Fig 14: Example 2

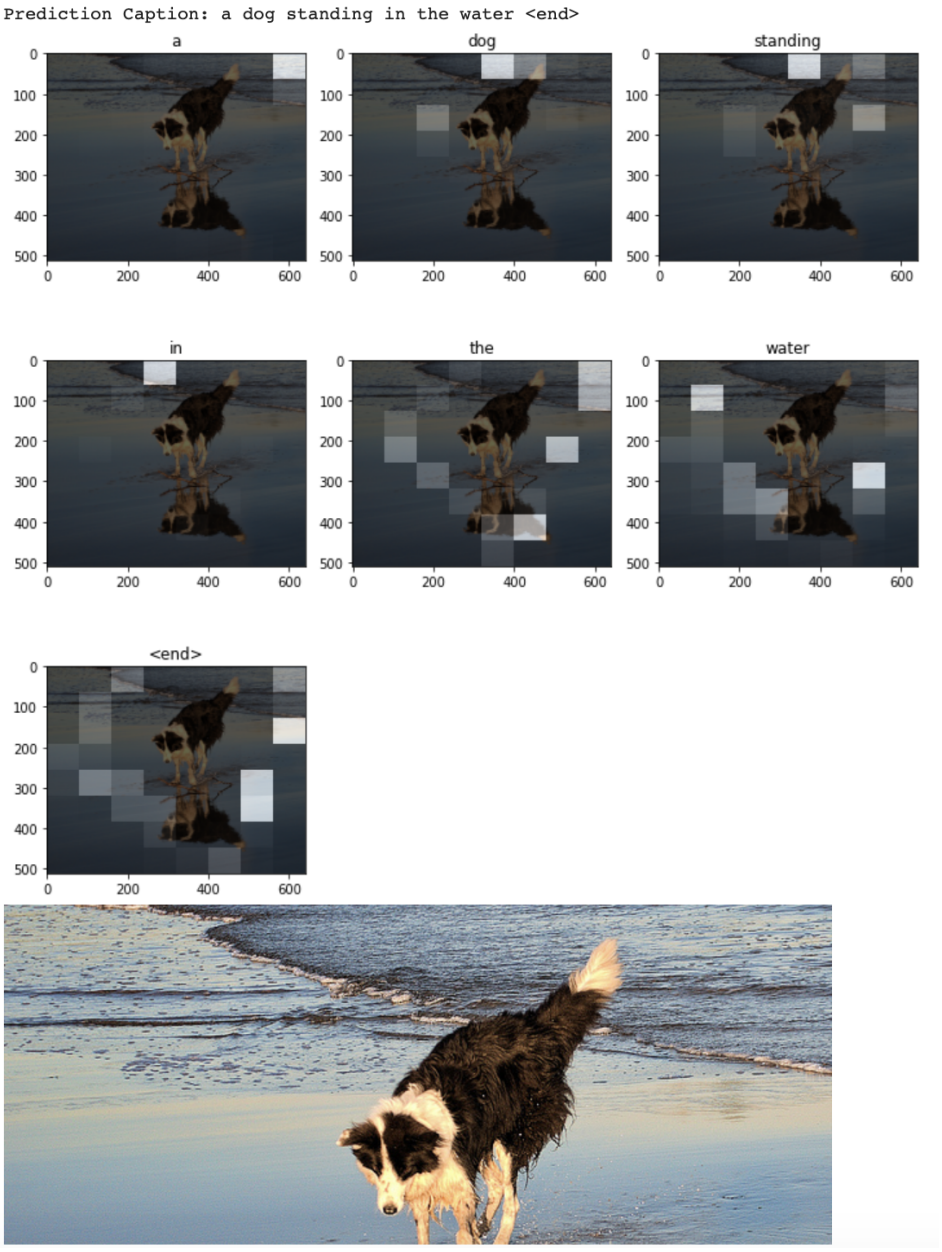


Fig 15: Example 3

Fig 16: Example 4

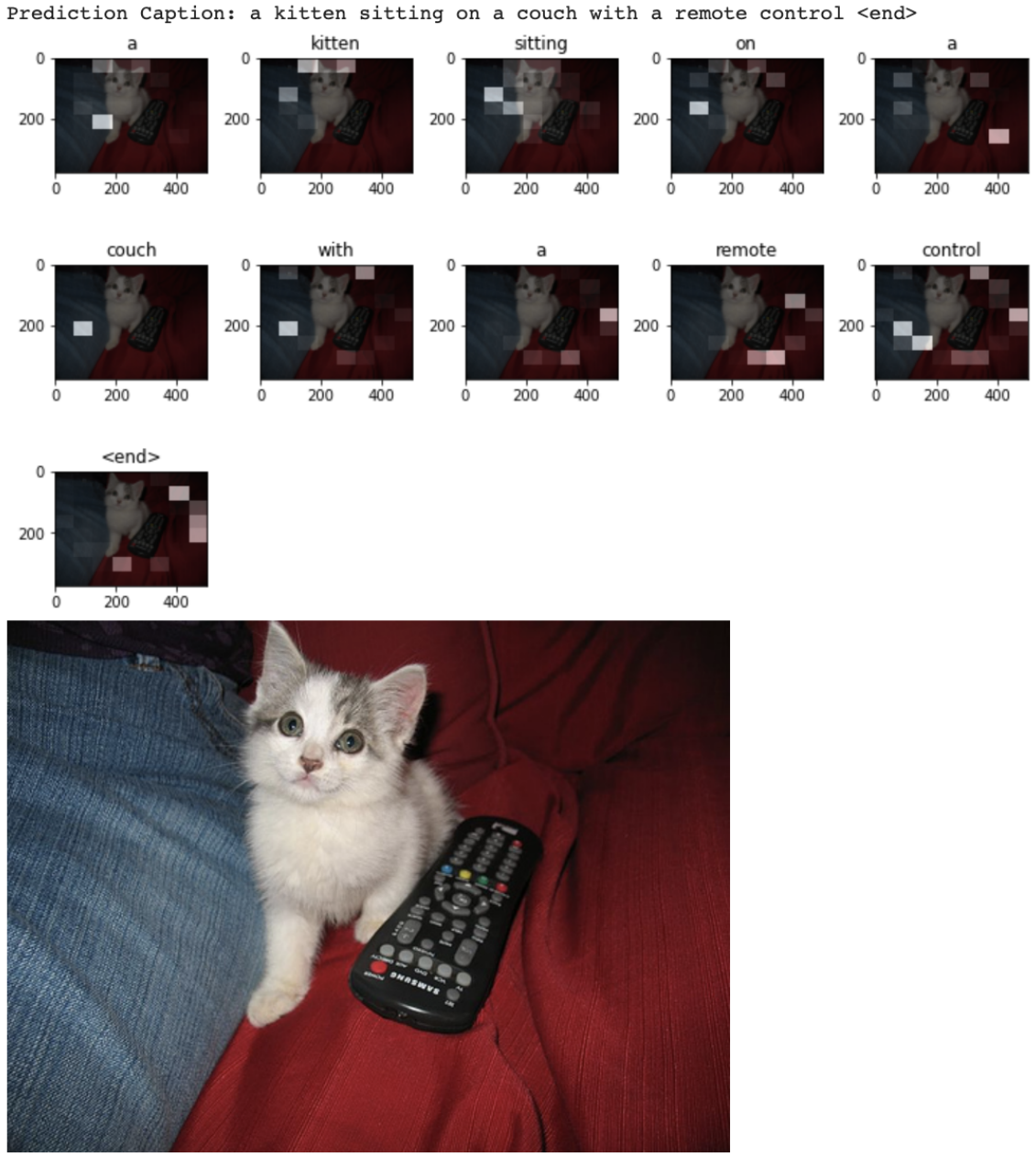


Fig 17: Example 5

CONCLUSIONS AND FUTURE ENHANCEMENTS

Chapter VIII

CONCLUSIONS AND FUTURE ENHANCEMENTS

VIII.I Conclusion

The deep learning model presented in this report uses Long-Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs) and successfully demonstrates an image caption generator tool. The image descriptions generated using the model described the images in detail with focussing on different objects in an image. It is also ensured that such a model is developed that is not biased or misguided on the basis of any one of the features present in the image that is being examined for caption generation. It was also seen that the trained model could detect motion of various objects in images as well as the orientations of those movements (*like moving up,sliding down,running etc.*). The presented model successfully extracts features from the input images and also they are described in natural language.

This report presents a comprehensive review of state-of-the-art work in the field of image processing,deep learning,computer vision,natural language processing and artificial intelligence. They made us better understand and adopt various techniques for feature extraction and also the evaluation methods for the model. The BLEU score is used in the presented model to evaluate it and a suitable accuracy value was achieved for the model. The state-of-the-art works also emphasised the need for improvement in the techniques used to describe an image automatically as the application domains of this model require critical use and scalability to fulfill its purpose. The software tools used in the development of the presented model have ensured that the large data set, which is a combination of images and their descriptions in a natural language, can be handled easily and effectively. Tools and packages like Keras, NumPy, Nltk package etc. provide various functions to ensure a smooth training and testing procedures while developing the deep learning model. After training the

model new mages are also described by the model with an acceptable accuracy level. Thus, making it a usable and scalable model for image generation tasks.

It is hoped that the effective implementation of the proposed machine learning model would eliminate many problems discovered during systems investigation.

VIII.II Future Scope

The facts to understand is that, although images have emerged to be a highly abundant data item across the globe and how it is highly important to analyze the image data, it is not the end. Videos are also available in huge amounts and they also carry the same amount of information and sometimes even more critical and useful information that when analyzed can be used in almost all the applications domain. They can sometimes have even more real time importance than describing the images especially for surveillance purposes. Hence, there is a need for a similar deep learning model that can describe the contents of a video in natural language.

Developing such a model would help in domains of security and military applications, real time crowd management, driverless vehicle technologies and helping visually impaired people etc. An extension of the present model can be developed to tackle this challenge of describing video contents in a natural language by a machine.

REFERENCES

1. Xu Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." International conference on machine learning. 2015.

2. Tanti, Marc, Albert Gatt, and Kenneth P. Camilleri. "What is the Role of Recurrent Neural Networks (RNNs) in an Image Caption Generator?." arXiv preprint arXiv:1708.02043 (2017).

3. Bernardi, Raffaella, et al. "Automatic description generation from images: A survey of models, datasets, and evaluation measures." Journal of Artificial Intelligence Research 55 (2016): 409-442.

4. Vinyals Oriol, et al. "Show and tell: A neural image caption generator." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.

5. Kuznetsova, Polina, et al. "Collective generation of natural image descriptions." Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. Association for Computational Linguistics, 2012.

6. Li Siming, et al. "Composing simple image descriptions using web-scale n-grams." Proceedings of the Fifteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2011.

7. Kiros Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models." International conference on machine learning. 2014.

8. Yang, Yezhou, et al. "Corpus-guided sentence generation of natural images." Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2011.

9. Zaremba, Wojciech, Ilya Sutskever, and Oriol Vinyals. "Recurrent neural network regularization." arXiv preprint arXiv:1409.2329 (2014).

10. Barnard Kobus, et al. "Matching words and pictures." Journal of machine learning research 3.Feb (2003): 1107-1135.

11. Yao Benjamin Z., et al. "I2t: Image parsing to text description." Proceedings of the IEEE 98.8 (2010): 1485-1508.

12. Kumar, N. Komal et al. “Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach.” 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS) (2019): 107-109.

13. Shabir, Sidra, and Syed Yasser Arafat. "An image conveys a message: A brief survey on image description generation." 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG). IEEE, 2018.

14. Li, Junnan, et al. "Video Storytelling: Textual Summaries for Events." IEEE Transactions on Multimedia (2019).

15. Li, Xiangyang, and Shuqiang Jiang. "Know more say less: Image captioning based on scene graphs." IEEE Transactions on Multimedia 21.8 (2019): 2117-2130.

JOURNAL DETAILS

Journal: International Journal of Computer Sciences and Engineering (e-ISSN: 2347-2693)

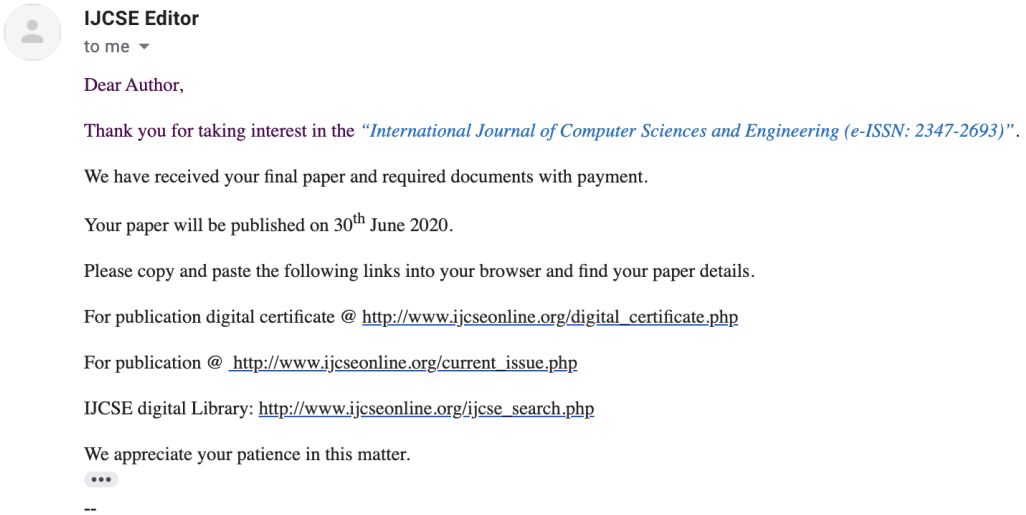
Paper: A Deep Learning Model for Image Caption Generator

Authors: P Aishwarya Naidu, Satvik Vats, Gehna Anand, Nalina V

Publication Details: Volume 8, Issue 6, June 2020

Link to journal: http://www.ijcseonline.org/

Paper Acceptance mail:



IJCSE Editor
to me

Dear Author,

Thank you for taking interest in the *"International Journal of Computer Sciences and Engineering (e-ISSN: 2347-2693)"*.

We have received your final paper and required documents with payment.

Your paper will be published on 30th June 2020.

Please copy and paste the following links into your browser and find your paper details.

For publication digital certificate @ http://www.ijcseonline.org/digital_certificate.php

For publication @ http://www.ijcseonline.org/current_issue.php

IJCSE digital Library: http://www.ijcseonline.org/ijcse_search.php

We appreciate your patience in this matter.

--

Current status: Paper accepted. Will be published on 30th June.

ankitha

ORIGINALITY REPORT

17%	11%	5%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	machinelearningmastery.com Internet Source	4%
2	Submitted to Banaras Hindu University Student Paper	2%
3	Submitted to Government College of Engineering Aurangabad, Maharashtra State, India Student Paper	1%
4	arxiv.org Internet Source	1%
5	www.citeulike.org Internet Source	1%
6	Submitted to Liverpool John Moores University Student Paper	1%
7	homepages.inf.ed.ac.uk Internet Source	1%
8	Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan. "Show and Tell: Lessons learned from the 2015 MSCOCO Image	1%

	Captioning Challenge", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016. Publication	
9	Submitted to CSU, San Jose State University Student Paper	<1%
10	Submitted to University of Wolverhampton Student Paper	<1%
11	Submitted to CSU, Dominguez Hills Student Paper	<1%
12	Xiangyang Li, Shuqiang Jiang. "Know More Say Less: Image Captioning Based on Scene Graphs", IEEE Transactions on Multimedia, 2019 Publication	<1%
13	Submitted to K. J. Somaiya College of Engineering Vidyavihar, Mumbai Student Paper	<1%
14	"Computing, Communication and Signal Processing", Springer Nature America, Inc, 2019 Publication	<1%
15	Submitted to Monash University Student Paper	<1%
16	Sidra Shabir, Syed Yasser Arafat. "An image	<1%

	conveys a message: A brief survey on image description generation", 2018 1st International Conference on Power, Energy and Smart Grid (ICPESG), 2018 Publication	
17	Submitted to University of Computer Studies Student Paper	<1%
18	Submitted to Kyungpook National University Student Paper	<1%
19	Huang, K.F.. "Heat exchanger network synthesis using a stagewise superstructure with non-isothermal mixing", Chemical Engineering Science, 20120507 Publication	<1%
20	Submitted to National Institute of Technology Uttarakhand Student Paper	<1%
21	worldwidescience.org Internet Source	<1%
22	Submitted to Asia Pacific Instutute of Information Technology Student Paper	<1%
23	Submitted to Cork Institute of Technology Student Paper	<1%
24	"Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series",	<1%

	Springer Science and Business Media LLC, 2019 Publication	
25	N. Komal Kumar, D. Vigneswari, A. Mohan, K. Laxman, J. Yuvaraj. "Detection and Recognition of Objects in Image Caption Generator System: A Deep Learning Approach", 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019 Publication	<1%
26	vnsgu.ac.in Internet Source	<1%
27	docplayer.net Internet Source	<1%
28	Armando Vieira, Bernardete Ribeiro. "Introduction to Deep Learning Business Applications for Developers", Springer Science and Business Media LLC, 2018 Publication	<1%
29	"Intelligent Technologies and Applications", Springer Science and Business Media LLC, 2019 Publication	<1%
30	upcommons.upc.edu Internet Source	<1%

31	lib.dr.iastate.edu Internet Source	<1%
32	www.mitpressjournals.org Internet Source	<1%
33	Jun Song, Siliang Tang, Jun Xiao, Fei Wu, Zhongfei Zhang. "LSTM-in-LSTM for generating long descriptions of images", Computational Visual Media, 2016 Publication	<1%
34	fiesta-iot.eu Internet Source	<1%
35	gujiuxiang.com Internet Source	<1%
36	Submitted to De Montfort University Student Paper	<1%
37	papyrus.bib.umontreal.ca Internet Source	<1%
38	MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga. "A Comprehensive Survey of Deep Learning for Image Captioning", ACM Computing Surveys, 2019 Publication	<1%
39	es.scribd.com Internet Source	<1%

40	diposit.ub.edu Internet Source	<1%
41	Submitted to University of Stirling Student Paper	<1%
42	www.mtome.com Internet Source	<1%
43	Submitted to Savitribai Phule Pune University Student Paper	<1%
44	link.springer.com Internet Source	<1%
45	"Data Management, Analytics and Innovation", Springer Science and Business Media LLC, 2020 Publication	<1%
46	Submitted to The University of Manchester Student Paper	<1%
47	Submitted to National Institute of Technology, Hamirpur Student Paper	<1%
48	Submitted to University of Limerick Student Paper	<1%
49	Submitted to VIT University Student Paper	<1%
50	Submitted to National College of Ireland	

Student Paper

<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		