

An Improved Technique for Effective Image Caption Generation

A THESIS REPORT

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE AWARD OF THE DEGREE
OF

MASTER OF TECHNOLOGY
IN
DATA SCIENCE

Submitted By
Ashish Verma (2K21/DSC/03)

Under the supervision of
Ms. PRIYA SINGH
(Assistant Professor, SE, DTU)



DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi – 110042

May, 2023

DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY

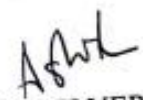
(Formerly Delhi College of Engineering)

Bawana Road, Delhi – 110042

CANDIDATE DECLARATION

I, Ashish Verma students of Master of Technology (Data Science), hereby declare that the project dissertation titled, "**An improved technique for effective image caption generation**" which is submitted by me to Department of Software Engineering, Delhi Technological University, Delhi in partial fulfilment and the award of the degree of Master of Technology, is original and not copied from any source without citation. It should be mentioned that it is original research work carried out by Ashish Verma during the academic year 2022-2023. This document has never been produced or formed the basis of evaluation for any previous diploma degree fellowship or any other title or recognition.

Place: DTU, Delhi
Date: May 2023


ASHISH VERMA
(2K21/DSC/03)

DECLARATION

Well hereby certify that the work which is presented in the Major Project-II/ Research Work entitled " An improved technique for effective image caption generation " in fulfilment of the requirement for the award of Degree of Master of Technology in Data Science and submitted to the Department of Software Engineering Delhi Technological University, Delhi is an authentic record of my/our own, carried out during a period from October , 2021, under the supervision of Ms. Priya Singh.

The matter presented in this report/thesis has not been submitted by us/me for the award of any other degree of this or any other Institute University. The work has been published/accepted/communicated in SCI/SCI expanded/SSCI/Scopus indexed journal OR peer reviewed Scopus indexed conference with the following details:

Title of the Paper: Empirical Validation of deep learning on image captioning models

Author names: Ashish Verma, Priya Singh

Name of Conference/Journal: 7th International Conference on Advanced Computing and Intelligent Engineering

Conference Dates with venue (if applicable): 23-24 December 2022.

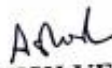
Have you registered for the conference (Yes/No)? Yes

Status of paper (Accepted/Published/Communicated): Accepted

Date of paper communication: Sept 26, 2022

Date of paper acceptance: Nov 15, 2022

Date of paper publication: NA


ASHISH VERMA
(2K21/DSC/03)

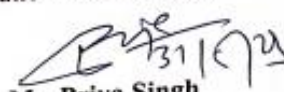
SUPERVISOR CERTIFICATE

To the best of my knowledge, the above work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere. I, further certify that the publication and indexing information given by the students is correct.

Date: May, 2023

Place: Delhi Technological University

Bawana Road, Delhi-110042


Ms. Priya Singh

SUPERVISOR
Assistant Professor, SE

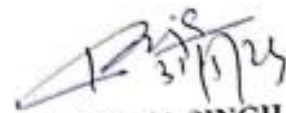
DEPARTMENT OF SOFTWARE ENGINEERING

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi – 110042

CERTIFICATE

This is to certify that the project entitled " An improved technique for effective image caption generation " prepared by Ashish Verma(2K21/DSC/03) students of Masters of Technology ("Data Science"), for the partial fulfilment of the requirements of the Masters of Technology degree, embodies the work, we all are doing during 4th semester of our course under due supervision of the supervisor from this college. It should be mentioned that it is original research work carried out by Ashish Verma(2K21/DSC/03) during the academic year 2022-2023. This work has not been produced or formed the basis of evaluation for any previous diploma degree fellowship or any other title or recognition.

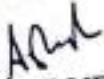
Place: DTU, Delhi
Date: May 2023


Ms. PRIYA SINGH
(Assistant Professor, SE, DTU)
SUPERVISOR

ACKNOWLEDGEMENT

I, Ashish Verma(2K21/DSC/03) would like to extend our gratitude towards Ms. Priya Singh for investing her valuable time and providing his guidance from time to time during the completion of this project without his support, guidance and recommendations it will be impossible to complete this project. We all want to sincerely thank our mentor for believing in us and boosting our confidence that made us do this project successfully and implement all the three shortlisted machine learning techniques and compare the results. We would also like to thank all the faculty members of Delhi Technological University, specifically the Department of Software Engineering who have helped us from time to time and provided us with their valuable guidance that helped us to complete our work within deadline.

Before proceeding with the overview of our project would also like to show gratitude towards all our friends, classmates and family and those helped us and given us required support in completing and finishing a project work.


ASHISH VERMA
2K21/DSC/03

ABSTRACT

An image is transformed into words through the practise of "image captioning". It is mostly employed in programmes that automatically require textual information in the form of data from each given image. These days, attention processes are extensively used in picture captioning models. In this case, the word generation may be distorted by the attention models.

In this work, we propose a Task-Adaptive Attention module to address this misleading problem in captioning pictures. This project performed over the two datasets, Flickr30k and MS COCO. BLEU, METEOR and CIDEr evaluated the target description sentence's likelihood given the training images. By contrasting the produced caption with the original caption, the BLEU score is determined

Content	Page Number
Candidate Declaration	ii
Declaration	iii
Certificate	iv
Acknowledgement	v
Abstract	vi
Index	vii
List of Figures	viii
List of Tables	ix
List of Abbreviations	x
CHAPTER 1 INTRODUCTION	1
1.1 Background	1
1.2 Motivation	4
1.3 Objective	4
1.4 Thesis structure	4
CHAPTER 2 LITERATURE REVIEW	5
CHAPTER 3 METHODOLOGY	9
3.1 Pre-processing steps	11
3.2 CNN model encoder	13
3.3 LSTM decoder	14
3.4 Adaptive attention model	16
3.5 Data Collection	19
3.6 Performance Measure	20
CHAPTER 4 EXPERIMENTAL ANALYSIS AND RESULTS	21
CHAPTER 5 CONCLUSION	27
CHAPTER 6 LIMITATION AND FUTURE WORK	28
REFERENCES	29
APPENDICES	
a. Plagiarism Report	42
b. Research Paper Acceptance	43
c. Proof of Registration	44
d. Scopus Index Proof	45

LIST OF TABLES

Table No	Table Name	Page _no
3.1	Image and captions corresponding to data points	12
4.1	Performance of hard attention and our adaptive attention model on Flickr30.	21
4.2	Performance of hard attention and our adaptive 23 attention model on COCO.	21

LIST OF FIGURES

Fig. No.	Fig. Captions	Page. no
1.1	This figure show attend and tell visualization	3
3.1	Basic System Architecture.	9
3.2	Proposed System Architecture	10
3.3	Stages of image captioning	13
3.4	VGG16 architecture	13
3.5	ResNet 50 Architecture	14
3.6	Inception V3 architecture	14
3.7	LSTM Architecture	15
3.8	ALSTM Architecture	15
3.9	Attention model	16
3.10	Proposed Adaptive Attention Model	17
3.11	Representation of a suggested model that, given a picture, generates the h1 target word X1.	18
3.12	Sample images of datasets along with caption.	20
4.1	Representation of captions and attention maps for pictures from the COCO and Flickr30k collections.	23
4.2	Representation of COCO bring out captions, visual grounding properties of each word produced by our model.	24
4.3	Representation of Flickr30k bring out captions, visual grounding properties of each word build by our model.	25
4.4	Probability plot on COCO.	25
4.5	Probability plot on flickr30k.	26

LIST OF ABBRIVATION

Abbrivate	Description
CNN	Convolutional Neural Network
NLP	Natural Language Processing
LSTM	Long Short Term Memory
BLEU	Bilingual Evaluation Understudy Score
METEOR	Metric for Evaluation of Translation with Explicit Ordering
ROUGE-L	Recall-Oriented Understudy for Gisting Evaluation
FNN	Feed Forward Neural
Atten	Attention
Proba	Probability

CHAPTER 1

INTRODUCTION

The process of generating the appropriate description from the scenes, activities, and objects is known as image captioning. Humans are good at highlighting and describing an image in great detail. The same task is a little more complicated when it comes to visual recognition models. Generating a meaningful caption for a visual is a tough job, but it may be very advantageous once completed. The ability to generate captions from photographs has various applications, ranging from helping the perceivably disabled to making it possible to automatically and affordably classify the millions of daily pictures uploaded to the Internet. Other uses for it include networking sites, driverless cars, CCTV cameras, video captioning, chatbots, proofreading software recommendations, and more.

1.1 Background

With the rapid advancement of digitization, there is an enormous volume of imagery and many linked texts [1]. Automated picture captioning has recently sparked a lot of research. The goal of automated picture captioning is to bring captions to explain the information of an image automatically. Automated image captioning significantly impacts online personal assistants, image indexing, editing program suggestion, and impaired assistance [2, 3]. Although describing an image is a simple task for a human, it is quite complex for a machine [4]. Captioning an image requires detecting the things in the image and capturing how these objects are connected.

Caption generation from the image is a more challenging task than image classification. The connections between the various items in an image must be considered while writing a visual description. In addition to visualizing the objects in the image, the features mentioned above need to be articulated in a language that others can understand. A linguistic model is needed that comprehends and conveys the image in natural language.

Image captioning aims to generate image text from the images. The majority of the scientific community on picture captioning has focused on captions with just one sentence. A single statement can only adequately describe a tiny piece of a picture in this style due to its limited rapid and dramatic. Krause et al. [3] introduced the Visual Genome corpus as the primary picture captioning collection (2016). When learnt on this

database, single powerful labelling models result in repetitive paragraphs that are unable to appropriately explain a variety of picture properties.

Natural language processing and computer vision researchers are becoming increasingly interested in picture captioning, commonly referred to as the automatic creation descriptions of images in natural environments. This description assignment is quite challenging since in addition to needing to recognize and describe any major things in an image, it also has to comprehend and produce sentences that are both syntactically and semantically accurate. The ability to generate captions from photographs has various applications, ranging from helping the perceivably disabled to make it possible to automatically and affordably classify the millions of daily pictures uploaded to the Internet. Other uses for it include networking sites, driverless cars, CCTV cameras, video captioning, chatbots, proofreading software recommendations, and more.

For paragraph generation, various approaches are employed, including Long-Term Recurrent Convolutional Network: An picture or series of images from a video frame can be used as the input. The concept is put into CNN, which detects activity inside the vision and produces a vector description of the image. The long short term memory model is then given this linear definition, which generates a term and a descriptive [4]. A communication system called an RNN is made to handle data with a records index t that is between 1 and t . Machine learning algorithms are frequently suggested for tasks that need sequential inputs, including speech development, such as those that require the prediction of the first word in a phrase.

Image captioning has demonstrated that visual attention is beneficial for a variety of vision-related activities, including categorizing, retrieval, and captioning of images and videos. It becomes a common method of using attention in deep learning networks for labeling images [29–32]. Most captioning models[20, 21, 22, 24, 23] have utilized the encoder-decoder structure, which was influenced by neural machine translation [25], and the CNN-based decoder produces the output after the RNN-based encoder has extracted visual information. Additionally, the attention mechanism was added to assist the model in concentrating on the pertinent places when creating each syllable [22, 14].

The model recommended using a CNN encoder to combine the two CNN models[5]. The pictures are encrypted using the VGG16, ResNet50, InceptionV3, and DenseNet model architecture. The CNN encoder collects characteristics from pictures and encodes them as readily understood numerical values. The recurrent neural network receives the retrieved

features (RNN). The caption is predicted in this study using LSTM and adaptive attention-based LSTM using the characteristics that were extracted.

The MS-COCO[35] and Flickr30k[36] datasets are used in this study's suggested work. Both Flickr30k and MS-COCO include 30k and 32k photos, respectively. There are 5 captions included for each picture in the collection. The flicker30k dataset was split into two separate folders: images and flicker text data. The captions are saved with distinct IDs for each image. Three sections make up the picture dataset: training, development, and testing. Eighty-five percent of the dataset is utilized to train the model, while fifteen percent is used to test the models. The caption is predicted by the model using the vocabulary. The performance evaluation of different CNN models was done using BLEU[34], METEOR[32], and CIDEr[33].

Irrespective of which phrase will be said next, the majority of attention models for visual question answering and picture captioning focus on the visual at each and every time step [31, 29, 17]. However, not every word in the caption corresponds to a matching visual cue.



Fig.1.1 This figure show attend and tell visualization.

1.2 Motivation

We also must grasp the significance of this issue from a practical perspective. Consider a few instances where finding a remedy for this issue would be extremely beneficial.

- Self-driving cars — Captioning the region around the car can aid the driverless system. Nowadays, everyone wants to move towards the era of driverless automobiles, which is one of the most difficult challenges.
- Aid to the blind---- The blind will be able to independently traverse the roads thanks to a gadget we can design. To do this, the scenario may be first translated into text, then speech, from the text. Since both Deep Learning applications have become more well-known.
- Surveillance CCTV are put all around modern society, and in addition to recording everything that happens, they may also create subtitles and sound an alert when something bad happens.
- Fully automated captioning may help Google Image Search catch up to Google Search in quality as every photo could be transformed into a description first, followed by a search based on the description.

1.3 Objective

- The main objective of the project is to describe the visual content in the form of text using image paragraph captioning by generating the textual description in terms of 4-5 sentences for the given input image using deep learning processing techniques.
- In this proposed work, we provide an innovative adaptive attention encoder-decoder system that offers the decoder a backup alternative. We also provide a fresh LSTM extension that generates an extra "visual sentinel."

1.4 Thesis Structure

Chapter 1 briefly takes you through the introduction of the topic thereafter section 1.2 shows the motivation behind the project, section 1.3 gives the objectives of the project. Chapter 2 takes you through the summary of previously related studies. Chapter 3 gives the complete information related to implementation & methodology, it also explains the different CNN models, different types of performance metrics and dataset description. Chapter 4 shows the experimental analysis and result that has been gathered and evaluation. Chapter 5 gives the conclusion on the complete idea of the topics and results; and in Chapter 6 discusses the future scopes and limitation related to the idea. Finally, some references that have been taken are mentioned.

CHAPTER 2

LITERATURE SURVEY

The encoder-decoder system, which was suggested for machine translation, is extensively used in the image captioning task is to consider translation from pictures to text. Most approaches used in the image captioning model have been divided into three essential parts: feature extractor, CNN model, and how the language model uses the image information.

The feature extractor compresses the input image into small size rendering that the other model components can use. The approach has been intensively researched in last years, with the growth of computer vision. At a previous time, CNN were widely employed [6], [7]–[9]. Regardless of the image's content, Anderson et al. [1] stated that the obtain features recorded by CNN subsequent to grid of equally sized image. To extract visual elements from observed salient picture patches, they suggested Faster RCNN based on bottom-up attention model. We also employ this strategy. To obtain improved picture characteristics, specific approaches [7], [13], [14] use GNN or self-attention to model the relationship between distinct image regions further. In this approach, we mimic this association through self-atten.

The automatic caption generator (ACG) plays a very integral part and has various trademarks (i.e., self-driving cars, traffic symbol classification, search using images, detecting the malicious activities in the background). Many alternative picture caption suggestion systems have been developed throughout the years.

The researchers have done a comparative study[5] on the image captioning model. They used two encoder models for comparison: InceptionV3 and VGG16, using the flicker8k dataset and the LSTM and beam search to decode the model. Researchers at AI produce high-level features of the image, and the AI lab employed a CNN for each conceivable object image. The best area matched to each phrase was determined using Multiple Instance Learning (MIL) [6]. On the MS- COCO dataset, this approach yielded a BLEU score of 22.9 percent.

The proposed model generates the next word using the visual features generated words and previously generated words as input. The two models uses the both feed forward networks, first employed for picture description creation in [8]. RNN is used instead of a FFN network because multimodal recurrent neural networks (m-RNN) allow for a variable length context [15]. Since then, [1] language models, [5], and [9] have come to be known as long short term memory and its variations. In order to help in caption synthesis, Jia et al. devised LSTM [16],

a variation of LSTM that adds semantic data extracted from an image. Recently, Transformer-based models for picture captioning have been developed [2], [7], drawing inspiration from the Transformer model's success in the field of machine translation in recent years [17]. In our comprehensive model, the Transformer architecture is also utilised. It is feasible to use latent stochastic variables to measure the data uncertainty. It enables the creation of several phrases that may each explain a video while accounting for different random elements.

The process of creating picture captions, which is often done through attention processes, depends on how the language model interprets the visual data. The language model's initial uses the single picture feature for once. At each time step, Soft Attention and Intricate Attention [5] concentrate on more important visual components linked to specific image areas. Several techniques [6], [15]-[18] construct captions by fusing nonvisual clues gleaned from last words with visual aspects as opposed to merely relying on observable data. The recommended Adaptive Attention, using [18] as an illustration, takes visual and sentence context information at each and every step. It effectively modifies their attention weights based on the phrase context and the global picture feature of an image. These approaches differ from ours in that the nonvisual quality in [6], [16], and [17] is generated from the previously created text. In contrast, the distinctively separate visual feature we utilise has a zero-start value, learns by back-propagation, collects helpful task-specific information, and has no relation to the phrase we previously created. This multi-modal conundrum may also be efficiently solved with the help of video captioning.

The automatic creation of captions for photos has become a well-known multidisciplinary research challenge in both academia and business. [22-27]. It can help users who are visually challenged and make it simple for users to browse and mobilize enormous volumes of often unstructured visual data. Fine-grained visual cues from the image must be included by the model in order to provide captions of the utmost quality. Recent research [25, 27, 28] has looked at neural encoder-decoder models associated with visual attention, where the attention mechanism typically creates an item emphasizing image regions relevant to each generated word.

The two main phases of the development of photo captioning are the traditional approach phase and the deep learning method phase [22, 21, 24, 14, 27]. In the early phases of the traditional approach, retrieval-based [5, 27, 7] and template-based [19, 26] techniques are two common ways to execute photo captioning. Retrieval-based techniques pull out one or a group of phrases

that are most comparable to a picture from a pre-specified sentence pool, in contrast to template-based techniques that generate slotted sentence templates and use detected visual concepts to fill in the slots.

The language model generates the following term using the previously produced words as well as the visual cues as input. The first time that feed-forward neural networks were employed for picture description generation was in [5], and both of the suggested multimodal log-bilinear models utilize them. In Multimodal Recurrent Neural Networks (m-RNN), recurrent neural networks (RNNs), which allow varying context duration, act as feed-forward neural networks [15]. LSTM and its variations are therefore frequently employed as language models [1, 2, and 3]. To include semantic data taken from the picture to direct the caption synthesis, Jia et al. introduced g_LSTM [4], an extension of LSTM. As a consequence of the Transformer model's [8] success in machine translation, Transformer-based model for picture translation have recently been developed. Based on the context of the sentence and the overall visual characteristic, the weights of the two components are effectively changed at each time step. The right consideration is given to both the visual and sentence text information. In contrast to the additive non-visual features used by [6], [7], and [8] methods, which draw their non-visual features from previously created sentences, the additional non-visual feature used by our method is initialized with 0's and learned using back-propagation to record useful task-specific hints.

Vinyals et al. [9] used CNN to encode the images, and RNN- CNN was used to decode the image's features into text. The researcher pre-owned the CNN model to build a new grouping of CNN around the image field. The researcher uses the LSTM and bidirectional LSTM for the textual description of the image and puts two models together via model embedding. Flickr8k, Flickr30k, and MS-COCO are used to obtain the best results.

Li et al. [17] uses glo-loc attention mechanism for image description. The researcher used the VGG16 CNN model for the image feature, Fast RCNN for object detection of the image, and Att-mech for glo and loc feature unification. ROUGE-L, CIDEr, METEOR, and BLEU performance measures are used for the empirical evaluation of results.

Yt et al. [18] proposed the attentive linear transformation for automatic image captions generation. The researcher uses the CNN model to extract the features of images and RNN for decoding the images. The researcher uses the benchmark dataset flicker8k and MS-COCO. Image captioning has demonstrated that visual attention is beneficial for a variety of vision-related activities, including categorizing, retrieval, and captioning of images and videos. It becomes a common method of using attention in deep learning networks for labeling images

[16–19]. Most captioning models[20-24] have utilized the encoder-decoder structure, which was influenced by neural machine translation [25], and the CNN-based decoder produces the output after the RNN-based encoder has extracted visual information. Additionally, the attention mechanism was added to assist the model in concentrating on the pertinent places when creating each syllable [22, 14].

The model recommended using a CNN encoder to combine the two CNN models[5]. The pictures are encrypted using the VGG16, ResNet50, InceptionV3, and DenseNet model architecture. The CNN encoder collects characteristics from pictures and encodes them as readily understood numerical values. The recurrent neural network receives the retrieved features (RNN). The caption is predicted in this study using LSTM and adaptive attention-based LSTM using the characteristics that were extracted.

CHAPTER 3 METHODOLOGY

In this research, we provide an innovative adaptive atten encoder-decoder system that offers the decoder a backup alternative. We also provide a fresh LSTM extension that generates an extra "visual sentinel." The proposed architecture of our model shows in the fig. 3.1.

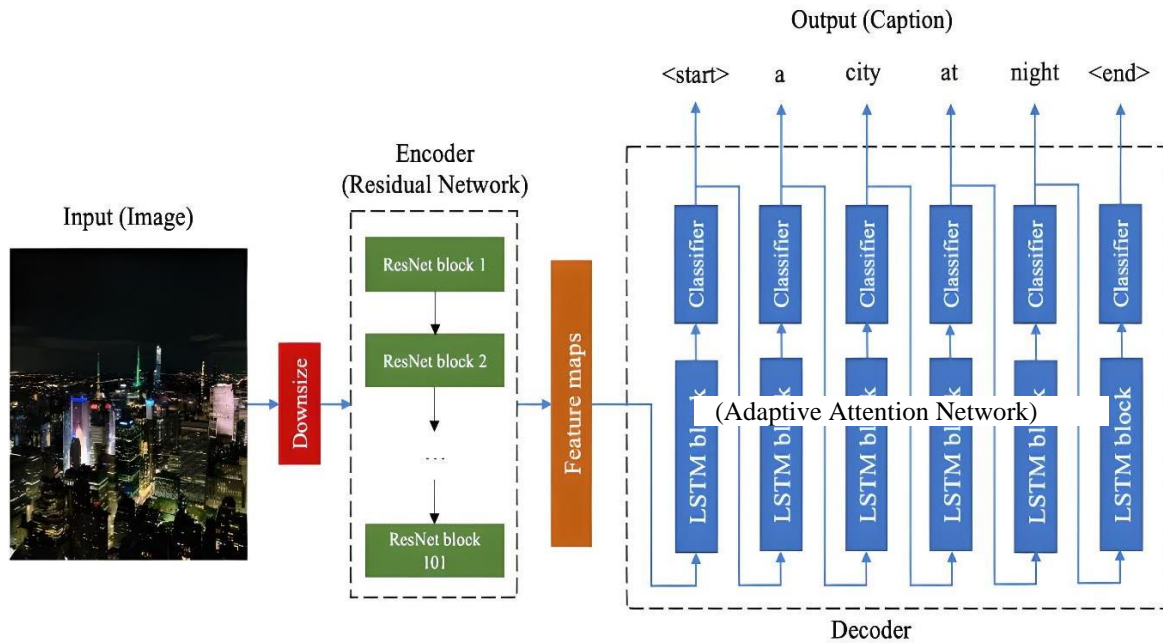


Fig 3.1 Basic System Architecture

The flicker30k and MS COCO dataset, which contains 30k and 328K images, is used in this study. The dataset is divided into train, test, and validate sets during data preparation for these photos.

Algorithm

- Step1. Preprocess the flicker8k and MS COCO dataset by downloading it.*
- Step2. Convert the text into tokens with spacy English tokenizer.*
- Step3. Adaptive attention is an object detector that may be used to extract image features.*
- Step4. Tokenization generates features on which CNN model is trained, and it generates captions.*
- Step5. All of the captions are combined to form a paragraph.*

In this section we have discuss about the system architecture.

PROPOSED SYSTEM

Image paragraph captioning is used to create descriptions from images. The text is generated using a hierarchical approach. Writing a caption for each object in the image is the first stage in the process. The final product is created by combining the captions.

Tokenization, that separates textual sequences into identifiers for use in data preparation, is the first component in this endeavour. The process is breaking apart a series of characters into components like words, phrases, symbols, and other things. The ids can be acquired from a file.

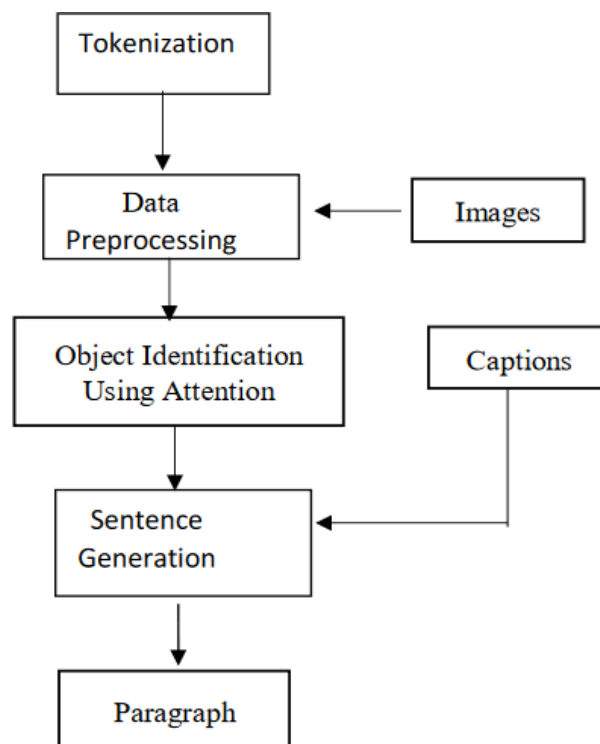


Fig 3.2 Proposed System Architecture

Preprocessing involves eliminating repetitions from data so as to retrieve it in its most basic form. In this instance, the dataset contains images that need to be enhanced. The dataset consists of three files with 14575,2489,2487 picture numbers each: train, test, and validate (image indices).

The final part of this effort is entity categorization [4,13], involves the detection of objects to facilitate the researcher's task. The Attention Model is used for the image captioning. The flow of execution is depicted in Figure 1. Initially, a picture is posted. Finding activity in the image is the first step. The recovered characteristics are then added to an Attention, which builds a

phrase by pulling a word from the physiological characteristic's vocabulary. After that, it advances to the Beginner level, when it generates a caption and many more sentences.

The final part in this project is sentence generation. Words are created simply employing source file characters as Captions and visual feature detection to identify things. Each word is combined with the previous word to form a phrase.

After 20 epochs, the CNN network is adjusted. We chose an 80-person batch size and trained for up to 50 epochs before halting early if the validation CIDEr [26] score did not rise during the previous six epochs.

3.1 Pre-processing steps

- **Data Preprocessing- Images**

Our model only accepts input (X) in the form of images. You might be conscious that an approach required all information to be delivered as a vector.

MS-COCO and Flickr8k datasets trained the model on 85% of the images, and the remaining 15% were used to test the model for each image having the five corresponding captions. The above-generated captions require some preprocessing to train our model. First, we need to load the features file containing all captions with the corresponding image IDs and loop the file, which maps each image with the corresponding IDs. Textual data should be filtered with discarding stop words like "a," "an," "the," and also token- containing digits. Then we need to create a vocabulary that contains all unique words across all image captions. To indicate the beginning and finish of captions, we must attach <startseq> at the beginning of the image caption and <endseq> at the end of the image caption.

It is necessary to convert each image into a set vector which that rnn can use.. To implement learning algorithm, we make use of the VGG16 (CNN) developed by Google Research. Using the flicker8k dataset, this model was trained to do image processing on 1000 distinct classes of pictures. Features extraction synthesis is the phrase for this; instead of identifying the image, For each, we want to get a fixed-length informative vector.

- **Data Preprocessing- Captions**

It's crucial to keep in mind that captions are something we should plan for. Consequently, throughout the training period, the model will be trained to identify captions as the data points (Y).

The full caption, however, is not foreseen at the same time as the photograph. We'll try to forecast the caption's terms. Every word must thus be encoded into a corrected vector. Two translations are Wordtoix (pronounced "word to index") and ixtoword (pronounced "index to word").

- **Data Preparation – Data Generation**

In the first example, Image 1, the black cat is described as "startseq the black cat sat on grass endseq". Always please remember that the source for forecasting is the picture vector and the caption. However, this is how we foresee the caption:

In ability to predict the syllable, i.e., we provide the picture vector and the first word as input.

In = Image1 + 'startseq'; Out = 'th'

third word, i.e.:Input = Image_1 + 'starteq the'; Output = 'ca'

The raw data matrix for one and associated captions can be made as follows:

Table 3.1: Image and captions corresponding to data points.

i	Xi		Yi
	Image feature vector	Partial Caption	Target word
1	Image_1	startseq	the
2	Image_1	startseq the	black
3	Image_1	startseq the black	cat
4	Image_1	startseq the black cat	sat
5	Image_1	startseq the black cat sat	on
6	Image_1	startseq the black cat sat on	grass
7	Image_1	startseq the black cat sat on grass	endseq

- **Feature Extraction**

The image serves as an input to the encoder model in our research. The visual data given for training the decoder is shared in fixed-size vectors. As a result, every image is turned into a fixed-size vector, then sustained into the RNN as input.

- **Model Training and Evaluation**

Train our model, we have used 1050 NVIDIA GPU. We have trained our decoder model on batch size 32 using the AdaOpt and catego_crostrophy as the loss function. In this study, the validation loss was determined as a metric to determine

the model's performance at each epoch. Finally, we discover the enhanced outcome, followed by the model being stored in a file.

3.2 CNN Model- Encoder

- **VGG16:** It consists of 16-layer networks for encoding the images. Out of 16 layers, 13 are the convolutional layer, and the rest are the dense layers. The VGG16 model architecture is expressed in Fig.3.4. The image dimension is $224 \times 224 \times 3$, used for feature extraction and the stride length is 1 for the CNN layer. The pooling layer employs the scaling factor of 2×2 pixels and a stride intensity of 2.

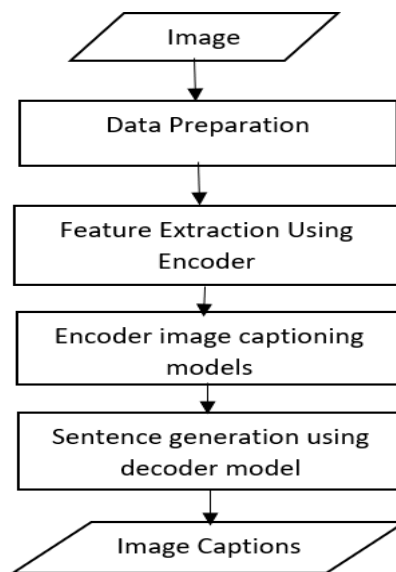


Fig.3.3 Stages of image captioning.

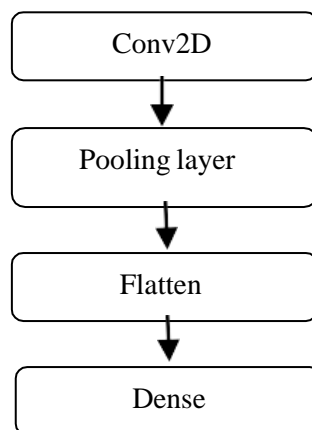


Fig 3.4 VGG16 architecture

- **ResNet50:** ResNet50 is the ResNet model, which consists of 48 convolutional layers along with one convolutional and one max pool layer. The Resnet50 architecture

expressed in Fig.3.5. The dimension of image used for feature extraction is 224*224.

- **InceptionV3:** The picture encoding task is performed by a 48- layer deep CNN in InceptionV3. InceptionV3 combines 11 inception modules with convolution layer and max-pooling layers in each. The image's dimension must be 229*229 in order for feature extraction to be performed on it.

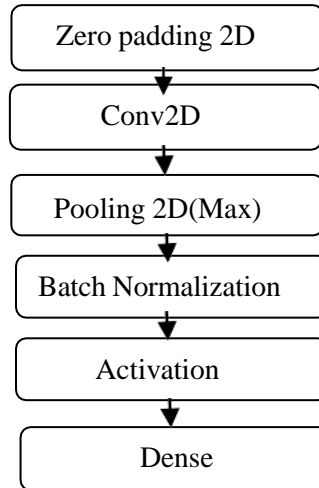


Fig3.5 ResNet50 architecture

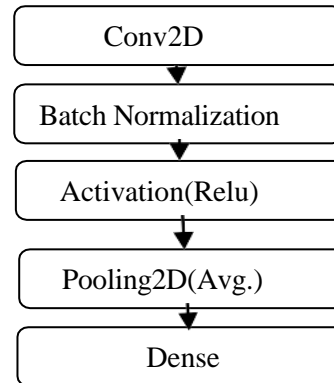


Fig 3.6 InceptionV3 architecture

3.3 Long Short Term Memory-DECODER

LSTM [8] is extensively used in the areas of audio-to-text conversion. LSTM primarily transports information from one cell to another and produces an entire word. LSTM consists of three gates: an input gate, an output gate, and a forget gate. A value is given to the input gate, which then sends it to other units. The forget gate determines how far a value will be used based on prior usage. All discounts are obtained through the output gate, which creates the output as a word. In addition to processing images, it can also process

data streams like speech or video.

LSTM is DL artificial RNN architecture. Feedback connections exist in the LSTM.

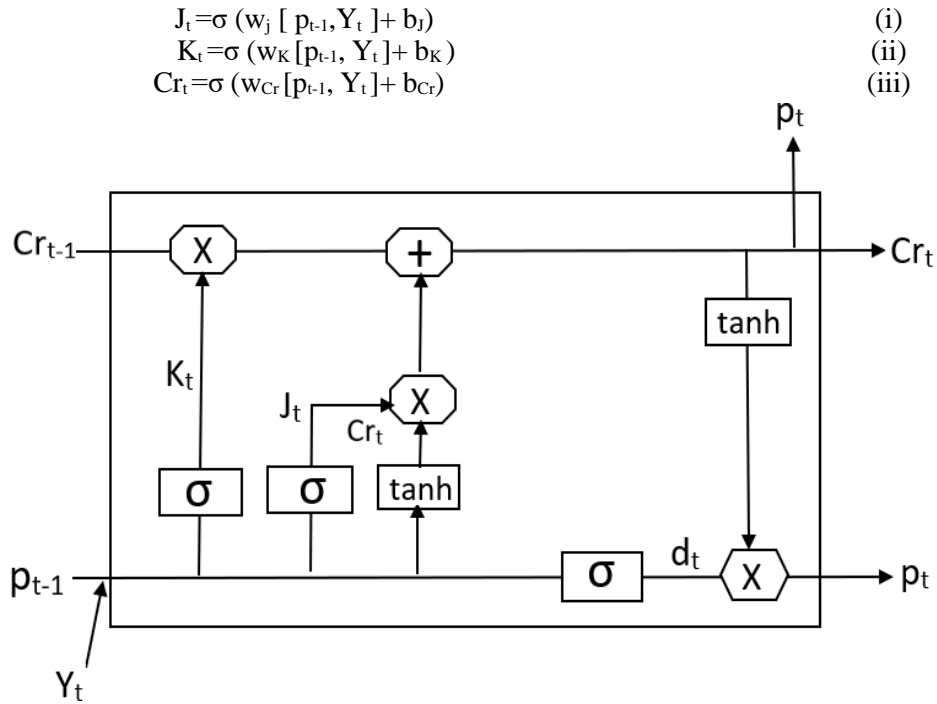


Fig 3.7 LSTM architecture.

It can manage selected data sequences (like images) and single data points (like utterances or visuals). A fundamental LSTM unit consists of a unit, an input gate, an output gate, and a forget gate. The three gates regulate the flow of information into and out of the unit, and the unit remembers values over an unbridled period. LSTM architecture is shown in Fig3.7.

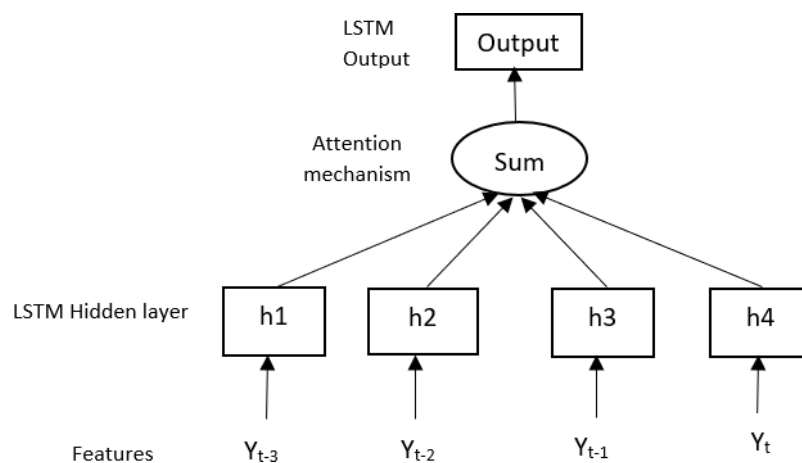


Fig 3.8 ALSTM architecture

The basic idea of the ALSTM model in the decoding phase, the attention weights of the ALSTM model, are used to choose pertinent encrypted concealed vectors from a series of instructive weights. At the time of creating the image descriptions, the ALSTM concentrates on the important areas of the image. The ALSTM architecture is shown in Fig3.8.

3.4 Adaptive Attention model- DECODER

Although geometric atten-based decoders models for captioning images have demonstrated effectiveness, they are unable to discriminate between situations in which they should rely on the input of visuals and those in which they should focus on the linguistic model. In this part, we put out a brand-new approach called "visual sentinel," which is an implicit presentation that the decoder brain actually knows. This concept was inspired by Merity et al. [19]. We expand our spatial attention model with "visual sentinel" and offer an adap atten model that may decide if it needs to attend the picture in order to forecast the following word.

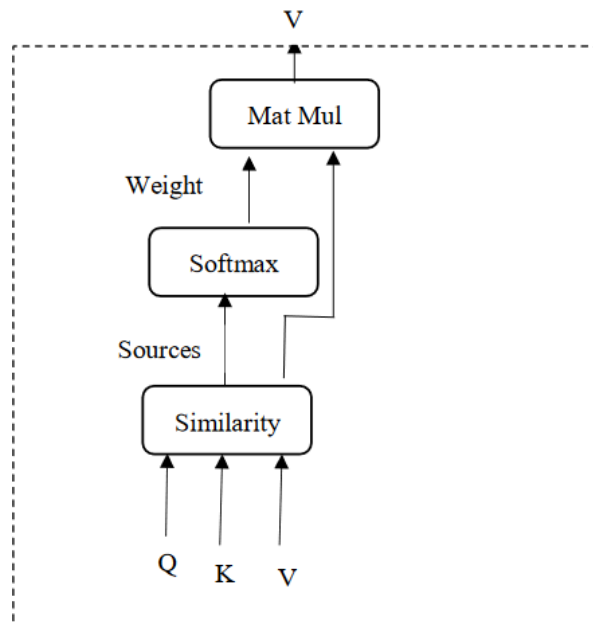


Fig.3.9 Attention Model.

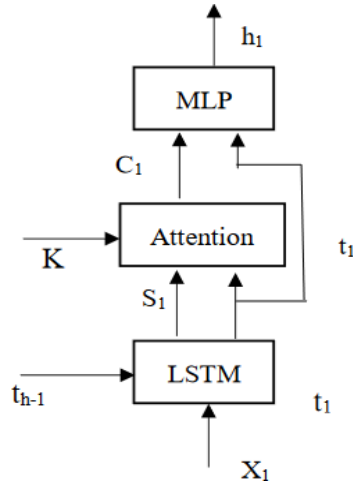


Fig.3.10 Proposed Adaptive Attention Model.

Fig.3.9 shows the architecture of the normal attention model and Fig.3.10 shows architecture of our proposed model. Adaptive attention is the most recent model object detection method based on Convolutional Neural Network. To improve the performance of previous architecture we had designed the different architecture. Adaptive attention model uses the multi-layer perceptron which can significantly detect the object more accurately as compare to the attention model. They found that the CNN performance benefits from carefully increasing the depth and width of their structures with respect to the spatial dimension. Some researchers have exploited the cardinality dimension. Others have found that skip and dense connections were also of benefit to performance. Recently, attention mechanisms on the channel dimension have gained popularity with researchers.

Couple long-term and short-term visual and language info is stored in the decoder memory. The model trains how to remove an additional aspect from this data when it chooses never to concentrate on the image. Visual sentinel is the name of this new element. The sentinel gate is the one that chooses whether to focus on the picture or the visual sentinel.

We consider the statistics kept in each memory block provided the decryption RNN is an LSTM. Thus, we extend the LSTM as follows in order to get the "visual sentinel" variable st .

$$z_t = sig(B_1 X_1 + B_h t_{h-1}) \quad (iv)$$

$$V_t = z_t \odot \tanh(mt) \quad (v)$$

where x_1 is the i/p to the LSTM, B_1 and B_h are weight variable that need to be learnt, and z_t is the gate put in to memory cell mt . Sig indicates logistic sigmoid activation, and \odot represents the elementwise product.

In order to create an appropriate vector related to the graphical sentinel, we build an adaptive atten strategy. Features of a geographically attended picture (i.e., vector of the location-based atten concept) and the visual sentinel vector are combined to create our innovative adaptive context vector, abbreviated as mt in our recommended design (see Fig. 4). By doing this, the network trades off shows much of the picture it considers as new information with it already knows in the decoder memory.

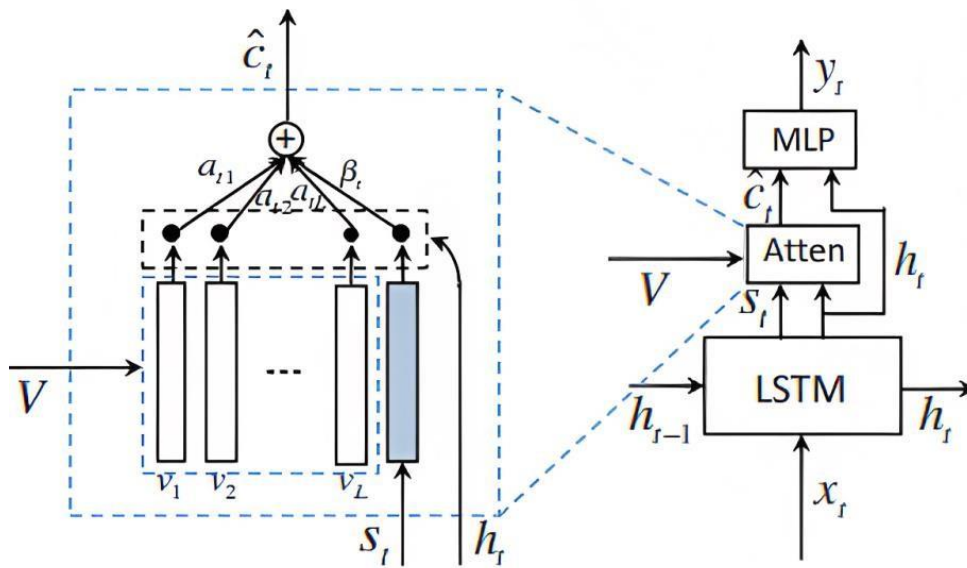


Fig.3.11 Representation of a suggested model that, given a picture, generates the h1 target word X1.

In order to create an appropriate vector related to the graphical sentinel, we build an adaptive atten strategy. Features of a geographically attended picture (i.e., vector of the location-based atten concept) and the visual sentinel vector are combined to create our innovative adaptive context vector, abbreviated as mt in our recommended design (see Fig. 4). By doing this, the network trades off shows much of the picture it considers as new information with it already knows in the decoder memory.

This is how mixing model is defined:

$$m^t = \beta t V t + (1 - \beta t) m^1 \quad (vi)$$

the new sentinel gate at time t, where t. T generates a scalar in our mixture model that falls between [0, 1]. When creating the following word, a value of 1 indicates that only the visual sentinel information is utilized, while a value of 0 indicates that only the spatial image information is used. We changed the spatial attention component in order to compute the

new sentinel gate t . We specifically add a new element to z , the vector that contains attention ratings as determined by Equation 6. This component reveals the degree to which the sentinel is receiving more "attention" from the network than its image characteristics. Equation 4 is changed to: to represent the inclusion of this extra ingredient.

$$\alpha^t = \text{softmax}(y^{t,:}, B_t h \tanh(B_s V_t + B_g h_t)) \quad (\text{vii})$$

where concatenation is indicated by $;$. The parameters for weight are B_s and B_g . Notably, the weight parameter in B_g is the same as the one in Equation 6. The attention distribution over 377, including the spatial image feature and the visual sentinel vector, is represented by α^t . The gate value is what we understand the last component of this vector to be:

$$\beta^t = \alpha^t [k + 1]. \quad (\text{viii})$$

One can compute the proba across a vocab of potential words at time t as follows:

$$j^t = \text{softmax}(B_p(m^t + h^t)). \quad (\text{ix})$$

where B_p are weight parameters.

This approach helps the propose approach to generate the next word while adaptively paying attention to the picture rather than the visual sentinel. At each time step, the sentinel vector is updated. We refer to our framework, which uses this adaptive attention model, as the adaptive encoder-decoder captioning of images framework.

3.5 Data Collection:

In the proposed work, this study used two datasets, flicker30k and MS-COCO, for the training and validation purpose of the model. Flicker30k[36] dataset provided by University of Illinois in Urbana-Champaign. This dataset consists of two folder images and text_data. Inside the image folder, there are 30k images, and every image has five corresponding captions. For each image, captions are stored with the IDs we have the unique image ids for each image. For training, the model dataset is splitted into three parts: training phase, development phase, and testing phase.

MS-COCO[35] dataset consists of 200000 images over the 330000 labeled. This dataset contains 80 object categories, the COCO classes, which include things for individual instance, maybe easily labeled person, car, chair, etc., and 91 stuff categories, including materials like the sky, street, grass, etc. MS COCO dataset also contains five captions per image.



The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.



A horse carrying a large load of hay and two people sitting on it.



Man in white shirt and blue shirt is standing in the camera

Fig.3.12 Sample images of datasets along with caption.

3.6 Performance Measures

Performance measures are used to evaluate the performance of different CNN models. In this noble we had used BLEU, CIDEr, METEOR for the performance evaluation of proposed model.

3.6.1 BLEU score:

A machine translations output is compared to a human translation using the BLEU[34] quality metre score. The fundamental tenet of BLEU is that the more an automation resembles a qualified human translation, the better. Always between 0 and 1, the BLEU score value lies. If the BLEU score is 0, generated captions are irrelevant to the actual captions. On the other side, if BLEU scores 1, generated captions are equal to the actual captions.

- Between the original translation and the derived interpretation, there is an N-gram disparity.
- Determine the accuracy for n-grams with sizes between 1 and 4.
- Make brevity punishable (for too short translation).

$$\min \left(1, \frac{\text{output}_{length}}{\text{reference}_{length}} \right)^4 \left(\prod_{i=1}^4 \text{precision}_i \right) \quad (\text{x})$$

- Usually calculated throughout the whole corpus, not just specific sentences.
- The BLEU measurement has ranging from zero to one.

$$\text{BLEU Score} = p \cdot e^{\sum_{n=1}^N \left(\frac{1}{N} \cdot \log P_n \right)} \quad (\text{xi})$$

- Only for a perfect fit, 1 is extremely uncommon.

3.6.2 METEOR: METEOR[33] is the metric used to evaluate computational linguistics output. The metric, which places more weight on recall than precision and is derived from the harmonic mean of unigram, precision, and recall, prioritises recollection. To combine accuracy and recall, the harmonic mean is used.

3.6.3 CIDEr: CIDEr[34] compares how closely the automatically generated captions resemble those written by humans. It can be used to solve the problem of the weak correlation between previous metrics and human judgment.

CHAPTER 4

EXPERIMENTAL ANALYSIS AND RESULTS

The result using the Flickr30k and MS-COCO caption assessment tool, which report the subsequent metrics : B-1, B-2, B-3, B-4, METEOR and CIDEr. Table shows the result on the Flickr30k and MS COCO.

Table.4.1 Performance of hard attention and our adaptive attention model on Flickr30.

Method	B-1	B-2	B-3	B-4	METEOR	CIDEr
Hard Attention	0.669	0.439	0.296	0.199	0.185	0.9875
Our Adaptive Attention	0.684	0.494	0.365	0.241	0.240	1.029

Table.4.2 Performance of hard attention and our adaptive attention model on COCO.

Method	B-1	B-2	B-3	B-4	METEOR	CIDEr
Hard Attention	0.718	0.504	0.357	0.250	0.230	0.9980
Our Adaptive Attention	0.741	0.592	0.445	0.335	0.278	1.087

Verifying the efficiency of the suggested framework involves comparing the whole model to abbreviated versions without a visual sentinel. Our adaptive attention model performs noticeably better than the spatial atten model, which raises the CIDEr value on Flickr30k and COCO from 0.9875/1.029 to 0.9980/1.085, respectively. We can see that our single model greatly passes all prior method in all metrics when compared to them. Our method advances the state-of-the-art on MS COCO from 0.250 to 0.335 for BLEU-4, 0.230 (MSM) to 0.278 for METEOR, and 0.9980 (MSM) to 1.087 for CIDEr. Similarly, our model remarkably surpass the state-of-the-art on Flickr30k.

The VGG16 and unidirectional LSTM model gives the B-1 score 0.5913 with batch size=32 on flickr30k dataset. But on increasing the batch size the B-1 score 0.6052. In case of MS COCO dataset the increase in the batch-size decreases the B-1 score.



a man riding a bike down a road next to a body of water.



an elephant standing next to rock wall.



a close up of a fire hydrant on a sidewalk.



a yellow plate topped with meat and broccoli.



a herd of sheep grazing on a lush green hillside.



a little girl sitting on a bench holding an umbrella.



a stainless steel oven in a kitchen with wood cabinets.



two birds sitting on top of a tree branch.

Fig.4.1 Representation of captions and attention maps for pictures from the COCO and Flickr30k collections. A relationship between attention areas and emphasized words is shown by different color.

We initially see the spatial atten weight for each word in the produced text in order to better comprehend our model. We just use bilinear interpolation to up sample the atten weight to the picture size (224 *224). Fig. 4 displays produced captions as well as the spatial atten maps for individual caption terms. The first two columns provide examples of success, and the last column has examples of failure. We can observe that our model picks up on alignments that closely resemble human intuition. Note that proposed model does look at legitimate components in the pictures, even when it generates wrong captions, it merely appears to be unable to calculate or identify textures and small categories. A longer selection of representation is provided in additive material.

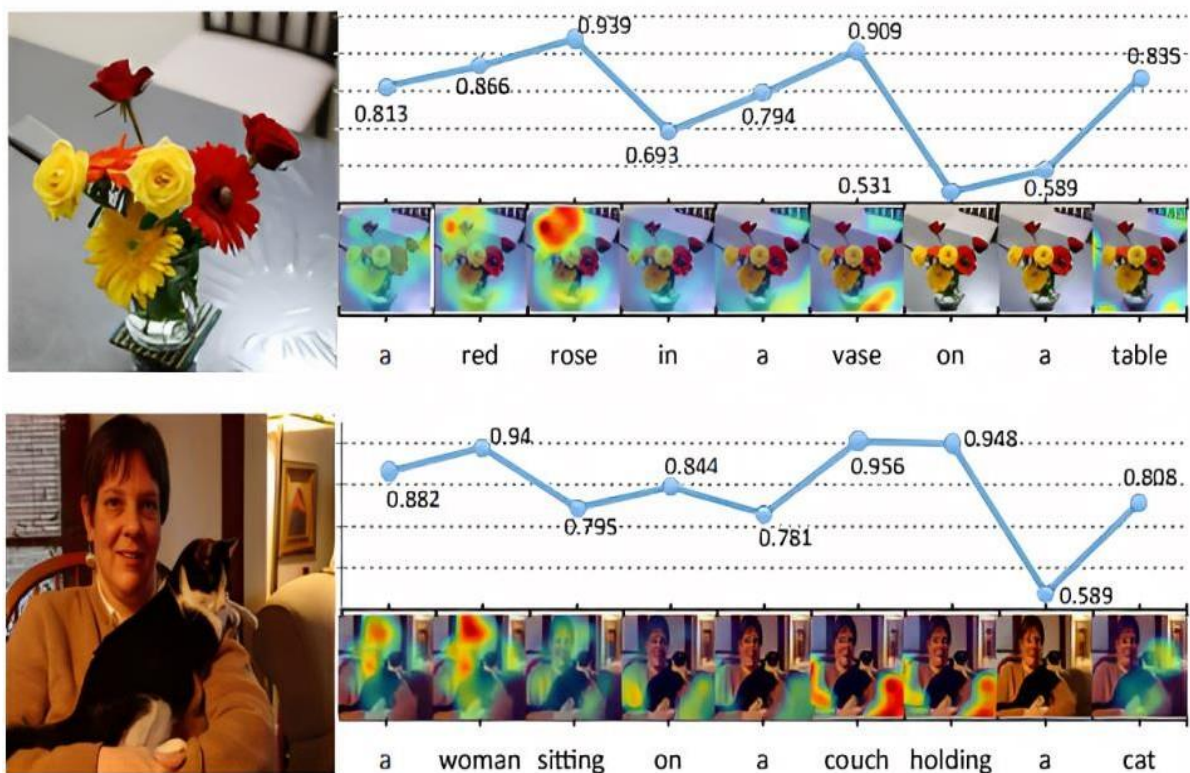


Fig4.2 Representation of COCO bring out captions, visual grounding properties of each word produced by our model.

As a caption is created, the sentinel gate is further visualised. We utilise the visual grounding probability of $1-\beta$ for each word. For each word in Fig. 5, we display the produced text, the proba, and the atten map. When producing non-visual terms like "of" and "a," our proposed model effectively learns to pay less atten to the image. Our approach gives words with strong visual connotations, such as "red," "rose," "doughnuts," "woman," and "snowboard," prob (above 0.9). Keep in mind that the same term may assigned different proba when cause in distinct factors.

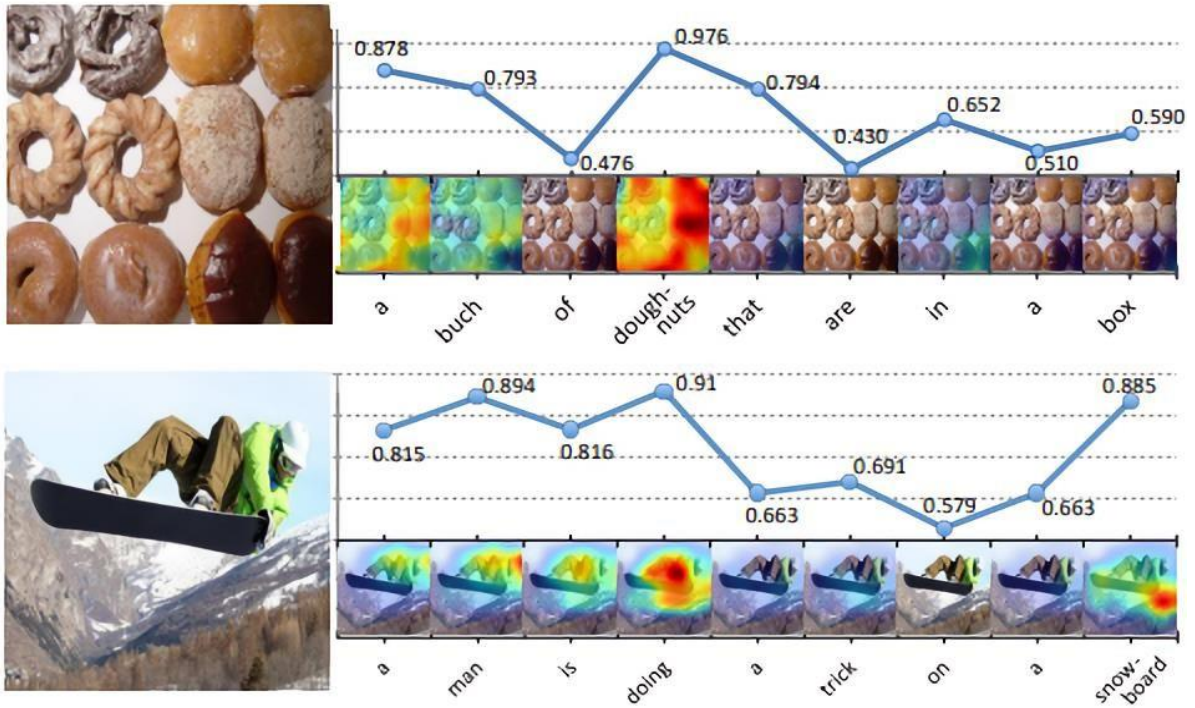


Fig. 4.3 Representation of Flickr30k bring out captions, visual grounding properties of each word build by our model.

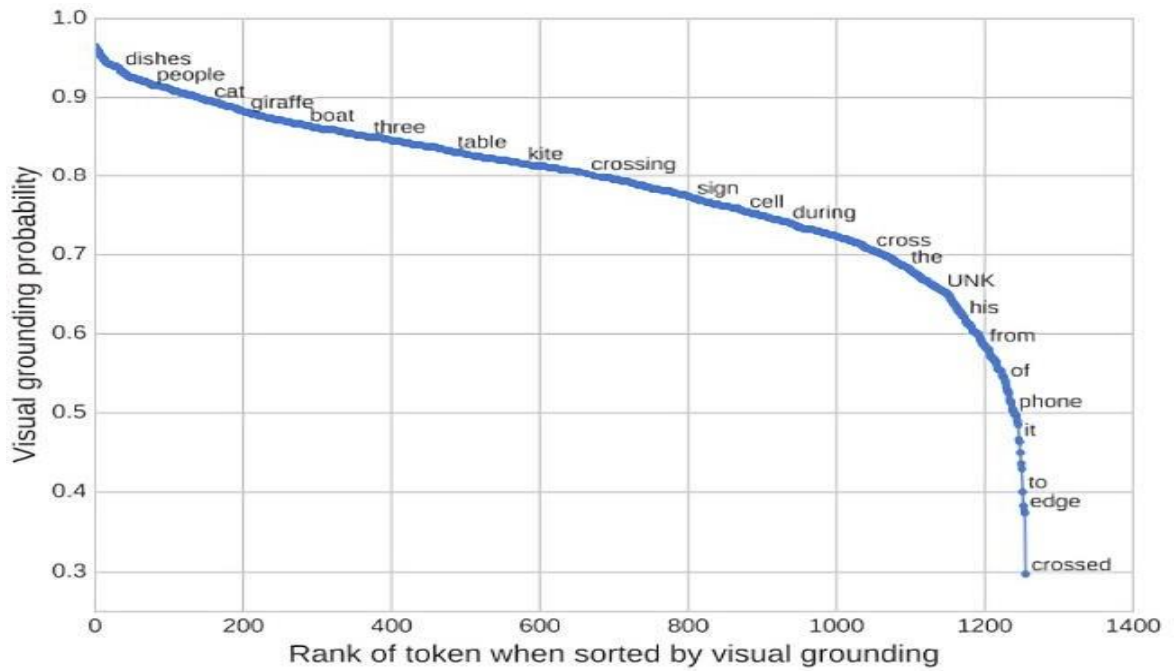


Fig.4.4 Probability Plot on COCO dataset.

For instance, the word "a" often has a excessive proba at the starting of a phrase since the model depends on the visual data to assess whether or not there is plurality without any language background. On the other hand, "a" in the phrase "on a table" has a far lesser chance of being visually grounded. The likelihood of something being on more than one table is low.

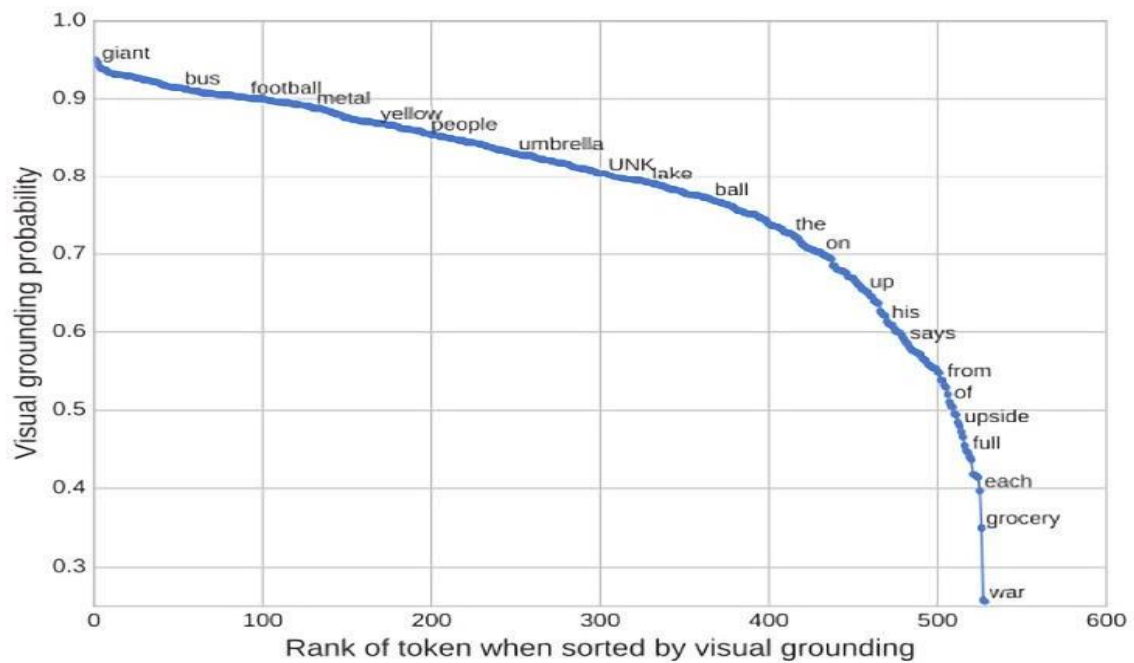


Fig.4.5 Probability plot on Flickr dataset.

CHAPTER 5 CONCLUSION

In this research, we provide an innovative adaptive atten encoder-decoder system that offers the decoder a backup alternative. We also provide a fresh LSTM extension that generates an extra "visual sentinel." Across common criteria for picture captioning, our model performs at the cutting edge. To analyze our adaptive attention, we do rigorous attention evaluation. Even though our model is tested on picture captioning, it may be used to other fields. The work's scope may be expanded in the future to allow all researchers to utilize the system more effectively. Implementing an Attention-based Model: The attention mechanism is becoming increasingly prominent. In the future, we may be able to utilize an updated attention-based different algorithms to focus on various sections of the image while the output sequence is being created. We will also use Hyperparameter Tuning in the future: The model's hyperparameters can be fine-tuned even more to improve the model's accuracy score.

CHAPTER 6 LIMITATION AND FUTURE WORK

Although experiments with specified models, datasets, and hyperparameters indicate promising results, the suggested study has some constraints, such as the lack of computers with higher processing power. If additional time is allowed, there may be possible improvements. First off, the network did not adapt to our particular training dataset since we employed a pre-trained CNN network straight as part of our workflow without any fine-tuning. In the future, as the output sequence is being constructed, we could be able to use an improved attention-based tool to concentrate on different areas of the image. We will also use Hyperparameter Tuning in the future: The model's hyperparameters can be fine-tuned even more to improve the model's accuracy score.

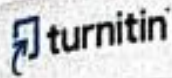
REFERENCES

1. P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, “*Bottom-up and top-down attention for image captioning and visual question answering*,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6077–6086.
2. S. Herdade, A. Kappeler, K. Boakye, and J. Soares, “*Image captioning: Transforming objects into words*,” arXiv preprint arXiv:1906.05963, 2019.
3. C. Yan, B. Gong, Y. Wei, and Y. Gao, “*Deep multi-view enhancement hashing for image retrieval*,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020
4. Q. Wang and A. B. Chan, “*Gated hierarchical attention for image captioning*,” in Asian Conference on Computer Vision. Springer, 2018, pp. 21–37.
5. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, “*Show, attend and tell: Neural image caption generation with visual attention*,” in International conference on machine learning, 2015, pp. 2048–2057
- 6 X. Yang, H. Zhang, and J. Cai, “*Learning to collocate neural modules for image captioning*,” arXiv preprint arXiv:1904.08608, 2019.
7. J. Yu, J. Li, Z. Yu, and Q. Huang, “*Multimodal transformer with multi-view visual representation for image captioning*,” arXiv preprint arXiv:1905.07841, 2019.
8. R. Kiros, R. Salakhutdinov, and R. Zemel, “*Multimodal neural language models*,” in International Conference on Machine Learning, 2014, pp. 595–603.
9. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “*Show and tell: A neural image caption generator*,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
10. C. Liu, H. Xie, Z. Zha, L. Yu, Z. Chen, and Y. Zhang, “*Bidirectional attention-recognition model for fine-grained object classification*,” IEEE Transactions on Multimedia, vol. 22, no. 7, pp. 1785–1795, 2020.
11. S. Min, H. Yao, H. Xie, Z. J. Zha, and Y. Zhang, “*Multi-objective matrix normalization for fine-grained visual recognition*,” IEEE Transactions on Image Processing, vol. 29, pp. 4996–5009, 2020.
12. J. Li, S. Zhang, and T. Huang, “*Multi-scale 3d convolution network for video based person re-identification*,” AAAI, vol. 33, pp. 8618–8625, 2019.
13. J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, “*Explain images with multimodal recurrent neural networks*,” arXiv preprint arXiv:1410.1090, 2014.

14. X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, “*Guiding the longshort term memory model for image caption generation*,” in Proceedings of the IEEE international conference on computer vision, 2015, pp. 2407–2415.
15. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “*Attention is all you need*,” in Advances in neural information processing systems, 2017, pp. 5998–6008.
16. . Lu, C. Xiong, D. Parikh, and R. Socher, “*Knowing when to look: Adaptive attention via a visual sentinel for image captioning*,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 375–383.
17. L. Gao, X. Li, J. Song, and H. T. Shen, “*Hierarchical lstms with adaptive attention for visual captioning*,” IEEE transactions on pattern analysis and machine intelligence, vol. 42, no. 5, pp. 1112–1131, 2019.
18. D. Wang and S. Zhang, “*Unsupervised person re-identification via multilabel classification*,” IEEE CVPR, 2020.
19. Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. “*Stacked attention networks for image question answering*”. In CVPR, 2016.
20. C. Xiong, S. Merity, and R. Socher. “*Dynamic memory networks for visual and textual question answering*”. In ICML, 2016.
21. J. Lu, J. Yang, D. Batra, and D. Parikh. “*Hierarchical question-image co-attention for visual question answering*”. In NIPS, 2016.
22. J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille. “*Deep captioning with multimodal recurrent neural networks (m-rnn)*”. In ICLR, 2015.
23. R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. “*Grounded compositional semantics for finding and describing images with sentences*”. 2014.
24. K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio. “*Show, attend and tell: Neural image caption generation with visual attention*”. In ICML, 2015.
25. H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. From “*captions to visual concepts and back*”. In CVPR, 2015.
26. A. Karpathy and L. Fei-Fei. “*Deep visual-semantic alignments for generating image descriptions*”. In CVPR, 2015
27. Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen. Encode, review, and decode: “*Reviewer module for caption generation*” In NIPS, 2016.
28. Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, “*Image captioning with semantic attention*” in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4651–4659.
29. L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T. Chua, SCA-CNN: “*Spatial and*

- channel-wise attention in convolutional networks for image captioning*”, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6298–6306.
30. X. He, Y. Yang, B. Shi, X. Bai, “*VD-SAN: visual-densely semantic attention network for image caption generation*” *Neurocomputing* 328 (2019) 48–55.
 31. J. Lu, C. Xiong, D. Parikh, R. Socher, “*Knowing when to look: Adaptive attention via a visual sentinel for image captioning*” in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3242–3250.
 32. METEOR-“*Metric for Evaluation of Translation with Explicit Ordering*”: <https://en.wikipedia.org/wiki/METEOR> (2023).
 33. CIDEr-“*Consensus-based Image Description Evaluation*”: https://www.oreilly.com/library/view/deep_learning (2023).
 34. BLEU-“*Bilingual understudy score*”: <https://en.wikipedia.org/wiki/BLEU>(2023).
 35. COCO- “*Common objects in context*”: <https://cocodataset.org/#home>(2023).
 36. Flickr30k- “*Flickr image dataset*”: <https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset>.

APPANDIX A- PLAGIARISM REPORT



Similarity Report ID: oia 27535-36509718

PAPER NAME
thesis.pdf

WORD COUNT
8740 Words

CHARACTER COUNT
46290 Characters

PAGE COUNT
45 Pages

FILE SIZE
2.0MB

SUBMISSION DATE
May 30, 2023 4:07 PM GMT+5:30

REPORT DATE
May 30, 2023 4:07 PM GMT+5:30

● 18% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 13% Internet database
- 7% Publications database
- Crossref database
- Crossref Posted Content database
- 11% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Quoted material

APPANDIX B- PROOF OF ACCEPTANCE

5/30/23, 11:06 AM

Gmail - NOTIFICATION: DECISION OF SUBMISSION PID-138: 7th ICACIE 2022, Springer



ASHISH VERMA <ashishverma01121999@gmail.com>

NOTIFICATION: DECISION OF SUBMISSION PID-138: 7th ICACIE 2022, Springer

1 message

ICACIE 2022 <icacie2022@easychair.org>

24 November 2022 at 21:22

To: Ashish Verma <ashishverma01121999@gmail.com>

Dear Ashish Verma, Priya Singh

Greetings from 7th International Conference on Advanced Computing and Intelligent Engineering.

Thank you for submission of your work in ICACIE 2022 and allow us time to review your work.

Congratulations!!!

We are happy to mention that your article entitled "EMPIRICAL VALIDATION OF DEEP LEARNING BASED ON IMAGE CAPTIONING MODELS" authored by "Ashish Verma, Priya Singh" is accepted for oral presentation and publication in Springer proceedings of 7th International Conference on Advanced Computing and Intelligent Engineering. The acceptance is provisional subject to timely submission of Camera Ready Article (less than 10% plagiarism), incorporating Reviewer's comment, and Registration.

Please visit <http://www.icacie.com> for author guidelines and prepare your article according to Springer template, failing to which Springer may not publish your article. The template and guidelines are given at <https://icacie.com/2022/author-guidelines/>

Kindly visit <https://icacie.com/2022/registration/> to properly learn the registration procedure and process of providing the camera ready paper. Incomplete registrations will not be considered. The submission of Camera ready paper along with all items [Camera ready version of paper (in PDF format), paper source file (Word Document/ Latex file), Publishing Agreement (Copyright Form), Permission Form, Payment Receipt, and Filled in Registration Form] in .zip / .rar format as mentioned in the Registration page need to be emailed to icacie2022@gmail.com on or before 30 NOV 2022.

If you need some more time to incorporate reviewers comments then please let us know but registration must be done on or before 26th Nov.

Looking forward to join with you during 23-24th DEC 2022 for conference at DRIEMS Autonomous Engineering College, Cuttack, Odisha, India.

Best wishes,

TPC Chair
7th ICACIE 2022

APPANDIX C- PROOF OF REGISTRATION



7th International Conference on
Advanced Computing & Intelligent Engineering
[ICACIE 2022]
23rd – 24th December 2022


No.: ICACIE2022/AUTHOR/138

Paper id: 138

MONEY RECEIPT

Received INR 7000/- with thanks from *Ashish Verma* towards Author Registration fees for the conference.

Author's Signature



Finance Chair

[Author's copy]

Organized by:
DRIEMS AUTONOMOUS COLLEGE, TANGI, CUTTACK, ODISHA, INDIA

APPANDIX D- PROOF OF SCOPUS INDEXING

7th International Conference on Advanced Computing and Intelligent Engineering

DRIEMS (Autonomous) Engineering College
Cuttack, Odisha, India
23-24 December 2022
www.icacie.com



Springer
ICACIE



UNIVERSITÉ DES
MASCAREIGNES
Esté - 1999

Call for Papers in Special Session

IMPORTANT DATES
Paper Submission: 30 SEP 2022
Notification of Decision: 15 NOV 2022
Author Registration & Submitting Camera Ready: 30 NOV 2022



UNIVERSITÉ DES
MASCAREIGNES
SAVOIR. C'EST POURTOI.

Proposals are invited for a Special Session organized by Université des Mascareignes, Pamplemousses, Mauritius in collaboration with DRIEMS (Autonomous) Engineering College, Cuttack, Odisha, India during the 7th International Conference on Advanced Computing and Intelligent Engineering (ICACIE 2022). This Special session will complement the regular technical program by highlighting new and emerging research topics or innovative applications of established approaches in areas such as Data Science, Big Data Analytics & Business Intelligence, which have dominated attention of Industry, commerce and research in recent years.

While we welcome special sessions in all areas of Advanced Computing & Intelligent Engineering, we are particularly interested in focused research topics, specifically those related to the following innovation themes:

Advances in Information System Artificial Intelligence Data Science Blockchain Technology Search Engine Analytics Machine Learning and Deep Learning Green Computing and Smart Cities Internet of Things and Edge Computing Cyber Security Cloud Computing Data Mining Big Data Analytics Recommender Systems	Robotics Intelligent Communication Systems Decision Science Data Driven Analytics & Business Management Financial Management Business Intelligence Marketing Analytics Human Resource Management Digital Humanities ICT in Education Risk Management Fintech and Digital Economy
--	--

Patron:
Radhakrishna Somanah CSK, Director General, Université des Mascareignes (UdM), Mauritius

General Chairs:
Bibudhendu Pati, RD Women's University, India
Hemant Kumar Rath, TCS Research & Innovation Labs, India

Special Session Chairs:
Seeven Amic, UdM Mauritius
Sanjeev K Cowlessur, UdM Mauritius

Special Session Committee Members:
Nirmal Kumar Betchoo, UdM Mauritius
Swaraj Horree, UdM Mauritius
Deojeet Nohur, UdM Mauritius
Swalehn Peeroo, UdM Mauritius
Kanayah Saurty, UdM Mauritius

Organizing Chairs:
Binod Kumar Pattanayak, S 'O' A DU, India
Chhabi Rani Panigrahi, RD Women's University, India

Website Chairs:
Subhashish D. Mohapatra, SoftASI Tech Sol., India
Sanjeev K Cowlessur, UdM, Mauritius

Publicity Chairs:
Coline Binmelix-Devalois, UdM Mauritius
Priyanta Ramtohol, UdM Mauritius

Advisory Committee:
Prasant Mohapatra, Univ. of California, Davis, USA
Rajkumar Buyya, Univ. of Melbourne, Australia
Sudip Misra, IIT Kharagpur, India
Laxmi N. Bhuyan, Univ. of California, Riverside, USA
Tomohiko Taniguchi, Fujitsu Labs Ltd, Japan
Chandan Samantary, Virginia State University, USA
Ashutosh Dutta, AT & T Lab, USA
Kuan-Ching Li, Providence University, Taiwan

All accepted papers of Special Session will be published in ICACIE 2022 proceeding, LNNS Series of Springer (Indexed in SCOPUS, Web of Science, Google Scholar etc.)

Publication Partners:



About Université des Mascareignes (UdM)

Université des Mascareignes (UdM) is the newest public university of the Republic of Mauritius set up in 2012 and aims to be the leading Higher Education Institution in the Indian Ocean. It has around 1100 local and international students and 60 academic staff distributed in three faculties:

- Faculty of Sustainable Development and Engineering
- Faculty of Business and Management
- Faculty of Information and Communication Technology

The UdM has three campuses across the country. UdM collaborates with Université de Limoges, France, to deliver double-degree awards. It follows the European Licence-Maîtrise-Doctorat (Bachelor, Masters, and PhD) system of higher education. UdM works in close collaboration with the Mauritian industry. All UdM programmes include a compulsory industrial internship component. Consequently, graduates of Université des Mascareignes have the unique potential of being highly employable and possess high skills to pursue further studies. In its endeavor to attain high academic standards, UdM collaborates with several international higher education institutions in the field of research and academia.

For more information, please visit:
www.icacie.com
Email: icacie2022@gmail.com

Paper Submission System:
<https://easychair.org/conferences/?conf=icacie2022>

