iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** I **Month of publication:** January 2021

DOI: https://doi.org/10.22214/ijraset.2021.32891

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue I Jan 2021- Available at www.ijraset.com

An Exploratory Analysis on Image captioning using Deep Learning

Anusha N G[1], H Chetana[2], Ashwini R[3], Meghana Reddy M[4], T H Sreenivas[5], Shruthi P[6]
[1, 2, 3, 4, 6]*UG Students, Department of CSE, Vidyavardhaka College of Engineering, Mysuru*
[5]*Assistant Professors, Department of CSE, Vidyavardhaka College of Engineering, Mysuru*

Abstract: ***Captioning an image is a concept of producing a succinct content description for an input image in single sentence considering all the objects in an image in the form of description. It can be done using deep learning architectures with the help of CNN (Convolution Neural Network) and RNN (Recurrent Neural Network). A particular kind of RNN called long short-term memory (LSTM) is used. The image from the dataset is taken as the input and accordingly caption is produced as an output from the given set in the form of text. It has numerous applications in various fields namely Image Indexing, Application Recommendation, Social media etc. It also aids the visually impaired and short sightedness people by automatically decoding the image and describing it in the form of text in a large format.***
Keywords: Deep Learning, CNN, RNN, LSTM.

I. INTRODUCTION

Deep learning is an intelligent retrieval function that emulates the functioning of human brain in various data transforming techniques. In the image caption generator model the image contents are automatically produced where computer vision and NLP (Natural Language Processing) are used. The model produces natural sentences which in time narrates the image. The CNN (Convolution Neural Network) and Recurrent Neural Network (RNN) are adopted. The LSTM (Long Short-Term Memory), a special kind of RNN is implemented that comprises of a memory cell, in order to keep and maintain the information for longer duration. However, the looming technologies in Deep Learning and the enormous progress in Natural Language Processing have equivalently improved the idea of captioning. Hence, newest appeals go after deep learning architectures which encrypt the visual features with Convolutional Neural Networks and decrypt using a language based model that converts the attributes and objects that were given by an image based model to a relevant sentence. The CNN is commonly made use to generate feature vectors using the temporal data in the images and thereby the vectors are then sent via the fully connected linear layer into the RNN architecture so as to produce the logical data or sequence of words which in the end produce illustration of an image.
In this paper, Section II shows an overview of the existing image captioning systems and Section III shows the comparison analysis performed among different methodology. Section IV is the conclusion of the paper.

II. LITERATURE SURVEY

In an automatic image description system [1], I2T and T2S systems are used to generate the description for an image and convert that description to speech form in order help the blind people. I2T system is used to generate the description in the text form, I2T system uses Residual Neural Network to extract image feature and bidirectional LSTMs to extract sentence feature and to generate description. T2S system is used to convert text form description to speech form. This system [1] uses Flickr8k dataset in which images of 6000 are utilized for teaching and the remaining images of 2000 are utilized for examining. The architecture of [1] is shown in the Fig. 1. BLEU (bilingual evaluation understudy) score is used to evaluate the captions generated. The BLEU score of this system [1] is 67.

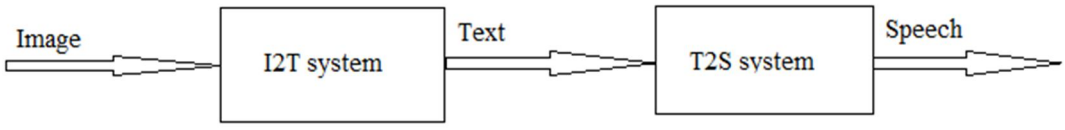


Fig. 1: Architecture of automatic image description system

Image captioning using deep neural architecture [2] uses show and Tell model to generate captions. By hybridizing two different models this model is created. An image is given as input to this model and then this image is given as an input to model of Inception v3. Last stage of the Inception-v3 model is made of a completely connected layer which converts the output of Inception-v3 model into a vector of word implanting. Vector of word implanting is given as input into LSTM serially arranged cells which help in storing and retrieving information which in a sequential order by time. This process helps in caption generation by keep tracking the previous words. Fig. 2 shows the model of show and tell architecture. In [2], MSCOCO dataset is used which contains 328k images. BLEU score of this system [2] is 65.50.

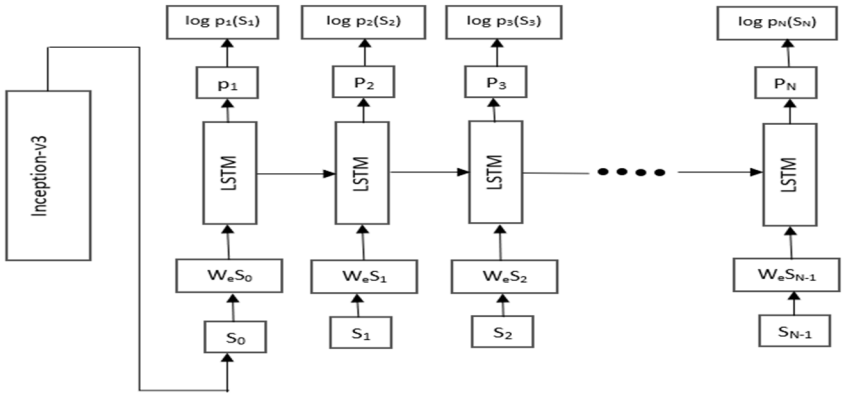


Fig. 2: Architecture of show and tell model

In an image caption generation system [3], CNN is used for encoding purpose and RNN (Recurrent Neural Network) is used for decoding purpose. These both are used to generate caption to an image. CNN is utilized to draw out the features of the image and RNN is utilized to generate the sentence.

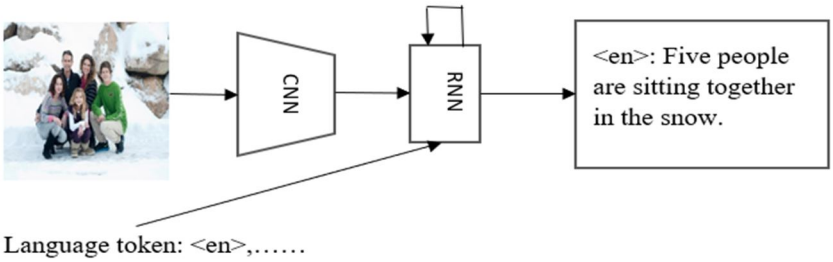


Fig 3: Model based on neural networks

In [4], CNN and RNN are utilized for caption generation to an image input. It uses Flickr 8k dataset to conduct experiments using python language to demonstrate the methods proposed. Fig. 4 shows the proposed methodology of [4].

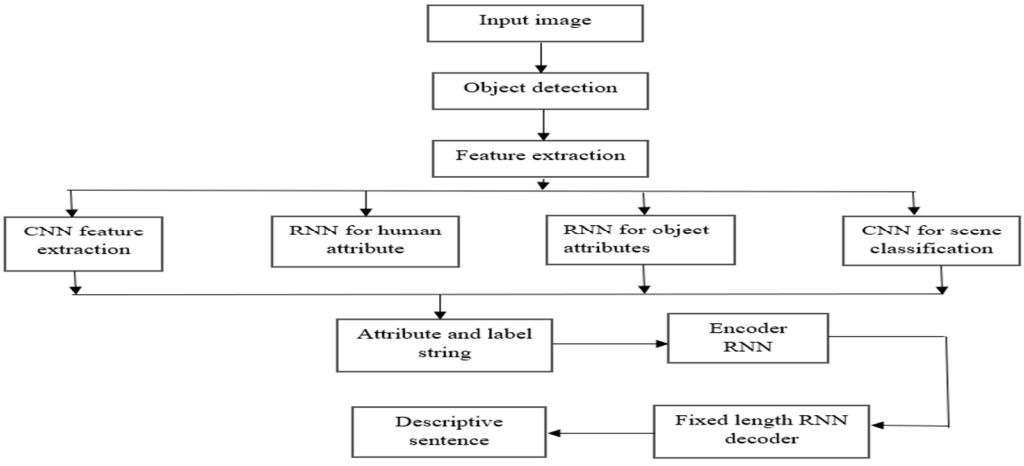


Fig. 4: Proposed methodology of [4]

In [5], Generative Adversarial Network (GAN) is used. It is a machine learning model which imagines the new samples where two models are trained at the same time. One is a generative model G which traps the data information, and the other is discriminative model D which determines the likelihood of how the sample shows up amid the training data preferable than G. In this model, a sample is extracted from the probability distribution among all the speculative images that tones with the interpretation.

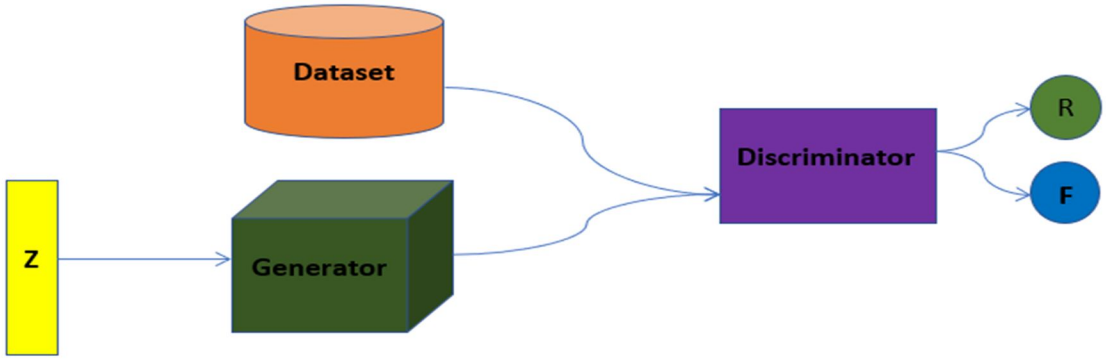


Fig. 5: Generative Adversarial Network Architecture [5]

The CNN and RNN are adopted in [6]. At first, the image is sent as input through the CNN to recognize the context of the image. And the transfer learning is used for the pre-processed model. The words are returned in the form of set as the output from the CNN model. Then, Natural Language Processing (NLP) is made use to interact with the system. At length, RNN is directed with the help of Flickr8k_text dataset. Correspondingly, the examined objects are sent to the RNN after certain intended processes and therefore RNN generates some relevant and significant caption.

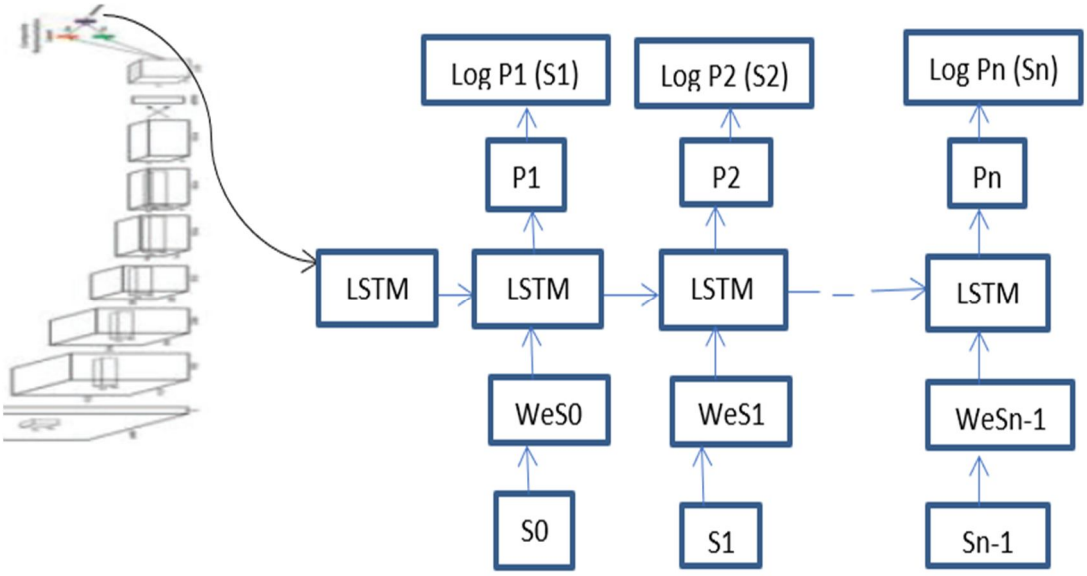


Fig. 6: Basic Workflow of the Model [6]

An encoder-decoder model in [7] is customized and optimized to work in a problem solving time and to function on mobile application devices. The Google's numerical computation library and the Tensor Flow are used in this model. The ductile construction of Tensor Flow allowed to install the computation model on CPUs and GPUs and consequently assisted to make use of an inherent parallelism in elementary functions and calculations. An end ProtoBuf file is produced that behaves as a Black Box of Image Captioning to set up a description for an image given as input and correspondingly, facilitating benefits, services. At last the redundant complications are extracted namely image reshaping, extraction of feature and forward is proceeded from the end-user or system.

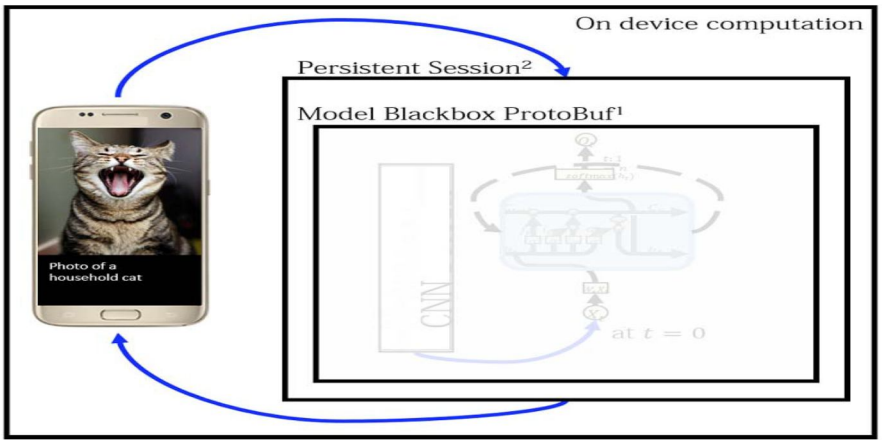


Fig. 7: Application Architecture

In Recognition and Detection of Objects in Generation of Image Caption System [8], important purpose here in paper is put forward the method of deep learning to generate caption utilizing neural networks. The Put forth methodology to generate captions as in perception also the identification as the objects utilizing deep learning as mentioned below Fig 8. This has the Pre-Processing of face Region Detection and Normalization, Handcrafted Feature Extraction, Deep Feature Extraction utilizing CNN Method followed by Image feature Concatenation, Image Feature Selection and Detection.

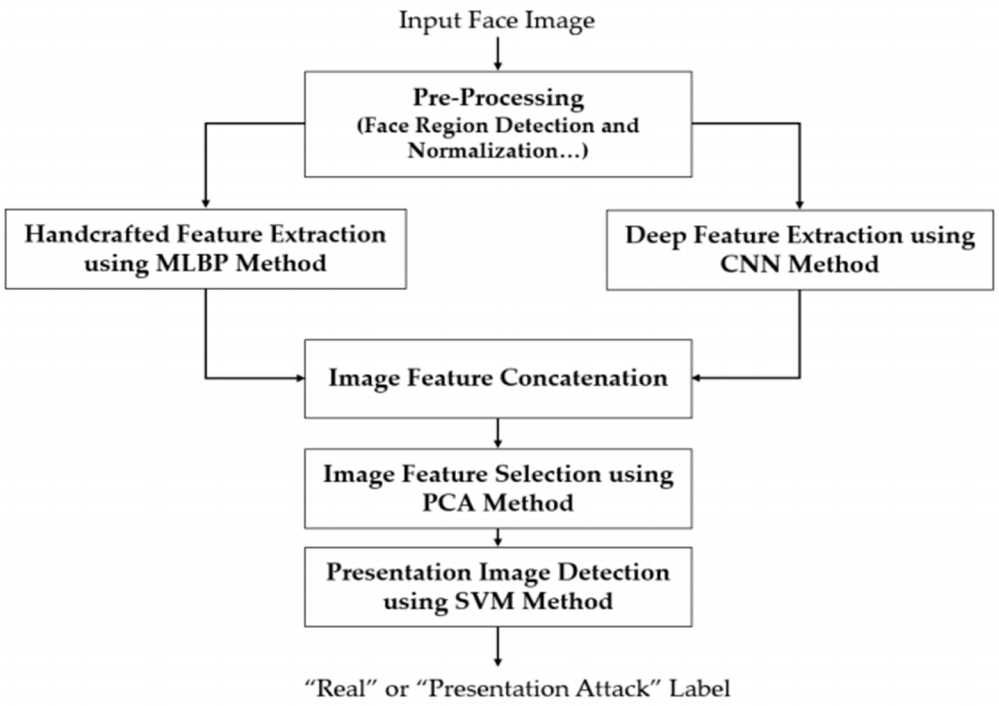


Fig. 8: Proposed Methodology for Generating captions

In the Black and White Images Captioned using Deep Learning Model [9], Machine learning method is used as one of the Transfer learning methods. Here, replica is developed to perform specific function which will be revised to be initiating position to the replica on the second task. Finding the solution to problem of captioning grayscale images without colorizing, at first flicker8k dataset is transformed into colorized images to grayscale using dot product of three channels considering the values for each pixel. Next we prepare images to be given as input to the Inception V3 model to extract features as it excepts only three channel inputs, for this we stacked up 1 channel grayscale image 3 times, then we normalize the images with standard deviation of 2 and mean 0.5. Then we input the image and extract the features.

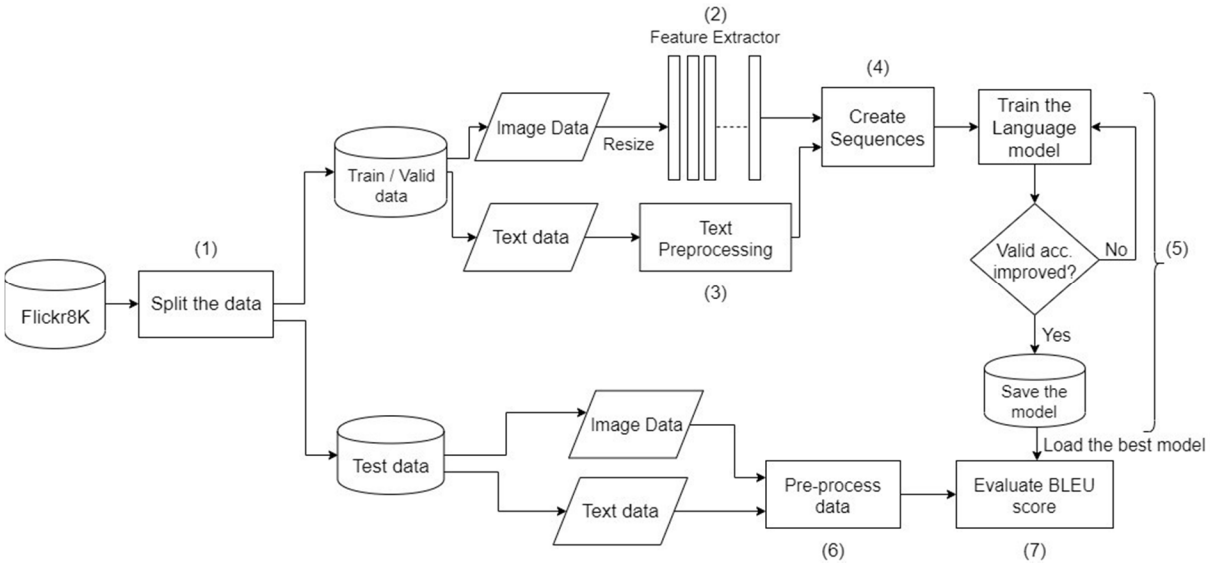


Fig. 9: Design of Dataset

Image caption generation of Cascade recurrent neural network [10] Captioning Image being the most accepted content in the computer sight, whose main purpose is to find correspondence between Visual and Language representation by briefing the contents of the images. As shown in Fig 10. the SGRU is developed in order of the different time scales which are adaptively captured dependencies by making each recurrent unit. It has two layers which are hidden consisting of gating units regulating the information within unit rather than supplying separate memory cell.

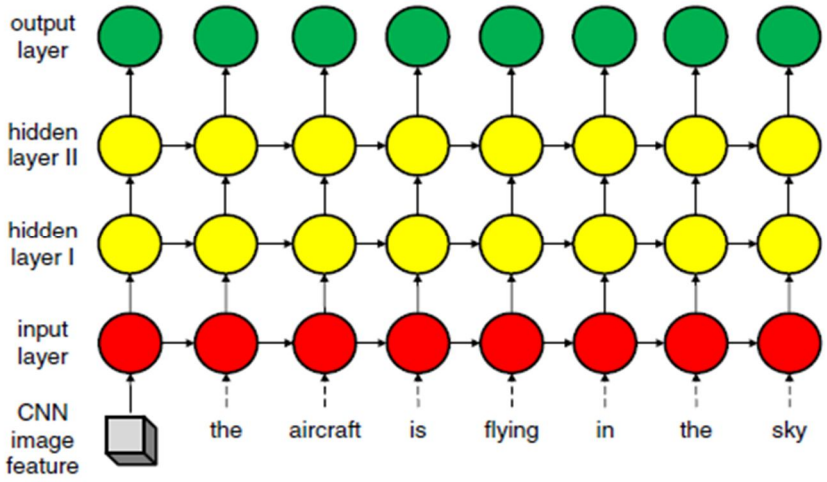


Fig. 10: Structure of SGRU.

In the above figure, red colour dot shows layer of input takes dense word representation. Yellow colour dot shows the hidden content with two integrated layers whereas the green colour dot represent layer of output.

Cascade Semantic Fusion for Image Captioning [11]. In this section, the details are based on the framework of encoder and decoder, where cascade semantic fusion architecture (CSF) is developed as a cipher, the structural features that encode the contents of an image are considered by default arranges for the attention mechanism, extraction of visual semantic features and also to combine to boost the performance of captioning. CSF works as of three cascade stages. As the first stage inculcates the contents of the objects, it basically depends on pertained detector. Second stage features fusion two times in the CSF. Initial characteristics of the combination intakes object as the input spatial characteristics. Hence image-level features along the context information. At third level, the spatial attention features are learnt to disclose the important section portrayal as the accompaniment of both preceding attention learned features. The spatial noticed module is adopted which induces the global features.

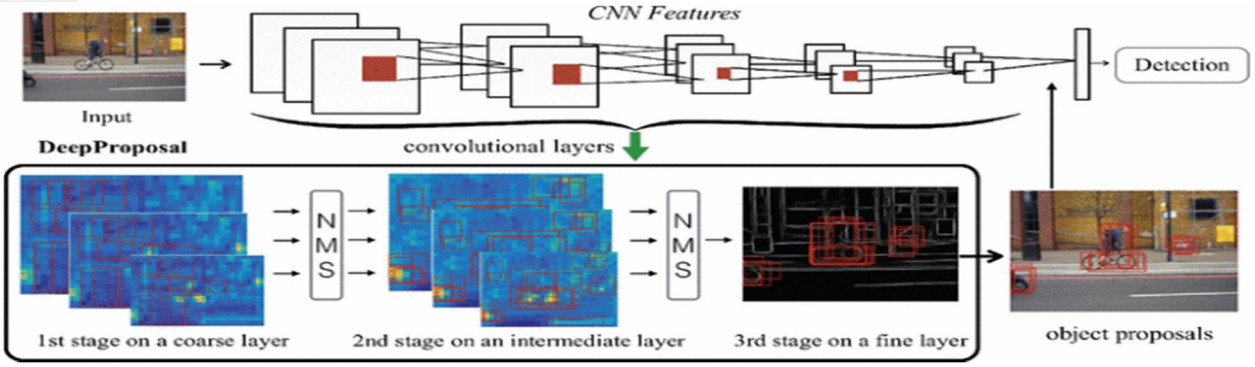


Fig. 11: Three stages of CSF Architecture

In the paper [12], the caption will be generated under mentioned content. These contents are grasped from the caption corpus that too by topic candidates. And to produce the caption to an image, an embedding vector is used to test the sector of the given image. And then language model is used to decode it into a sentence as output.

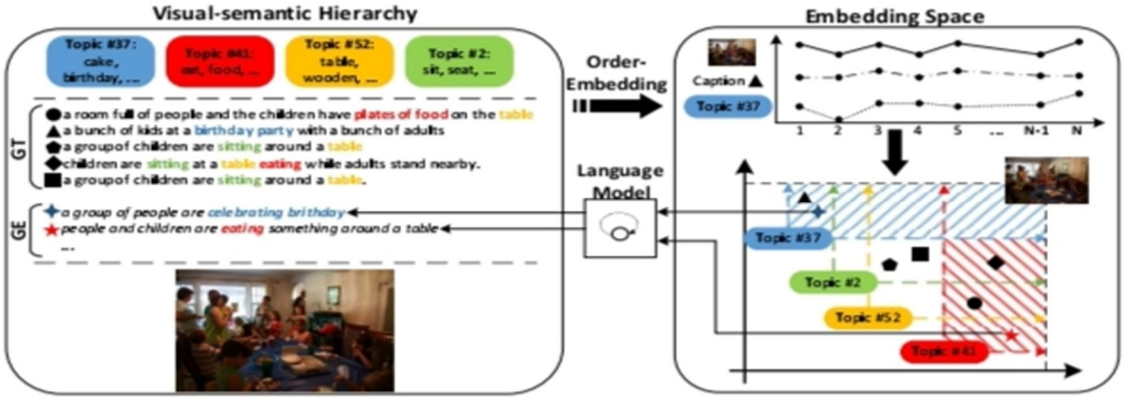


Fig. 12: Pipeline of the proposed method [12].

In [13], a network of sequential guiding used the proposed method of the pipeline. There is five ground-truth and four topics caption on the right of a single image and here it is plotted from the order-embedding into the N-dimensional embedding space. In the top of right corner, the hierarchical relationship is maintained in every coordinate. The vertical axis here refer to the value of every coordinate. Whenever, a caption is produced for an image, a point is tested in a sector of the embedding space. Then, a sentence will be generated with the help of language model by which it will be decoded.

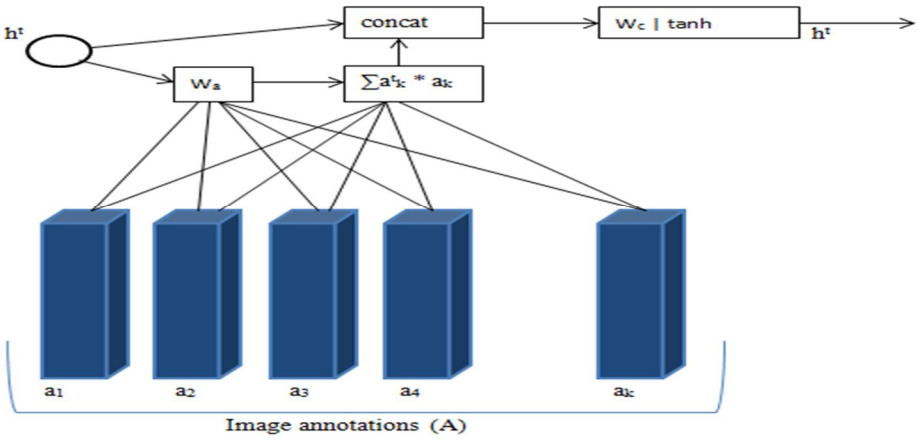


Fig. 13: An attention of Luong executed on image region.

In an Automated Image Captioning [14], the network's architecture contains a component called attention which takes the control of quantifying and managing the interconnection between the input and output elements named as General Attention and inside the input elements named as Self-attention. Loung attention is utilized by the top hidden layer states in both of encoder and decoder. The ResNet101 model pertained on image net is used as encoder. A photograph input is read by network model and the content is encoded. The LSTM network is utilized by decoder.

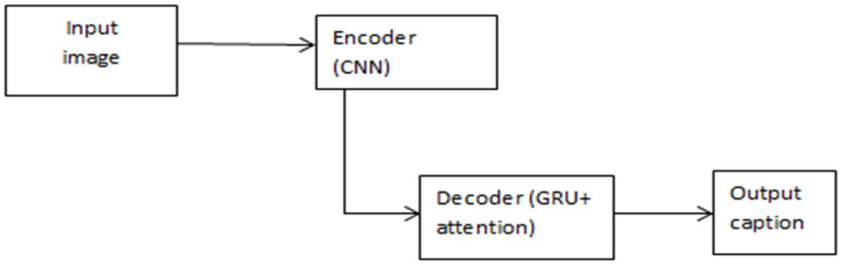


Fig. 14: Encoder-Decoder High-Level architecture.

Remote sensing Image Captioning [15], here the multimodal and the Attention-Based methods are used. These methods will use neural network to produce caption for remote sensing image. Represents remote sensing images, representing sentence and sentence generation are the three methods will be used. Attention-based method will use the two different types of manners, one is problematic manner and the other one is fatalism manner. Problematic manner will instruct the model by increasing the lower bound while the fatalism manner users will instruct by using back propagation techniques.

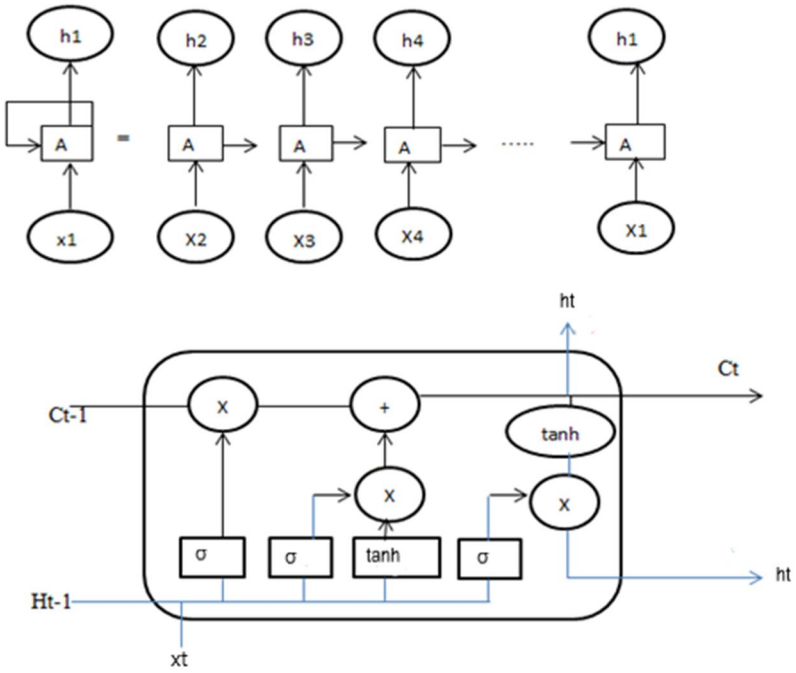


Fig. 15: Structure of LSTM

III. COMPARISON ANALYSIS

In [4], the experiment is done using different datasets and the evaluated using BLEU score. Table I. shows the BLEU score of different datasets.

Dataset	BLEU score
Flickr8k	0.53356
Flickr30k	0.61433
MSCOCO	0.67257

Table I. BLEU score of different datasets

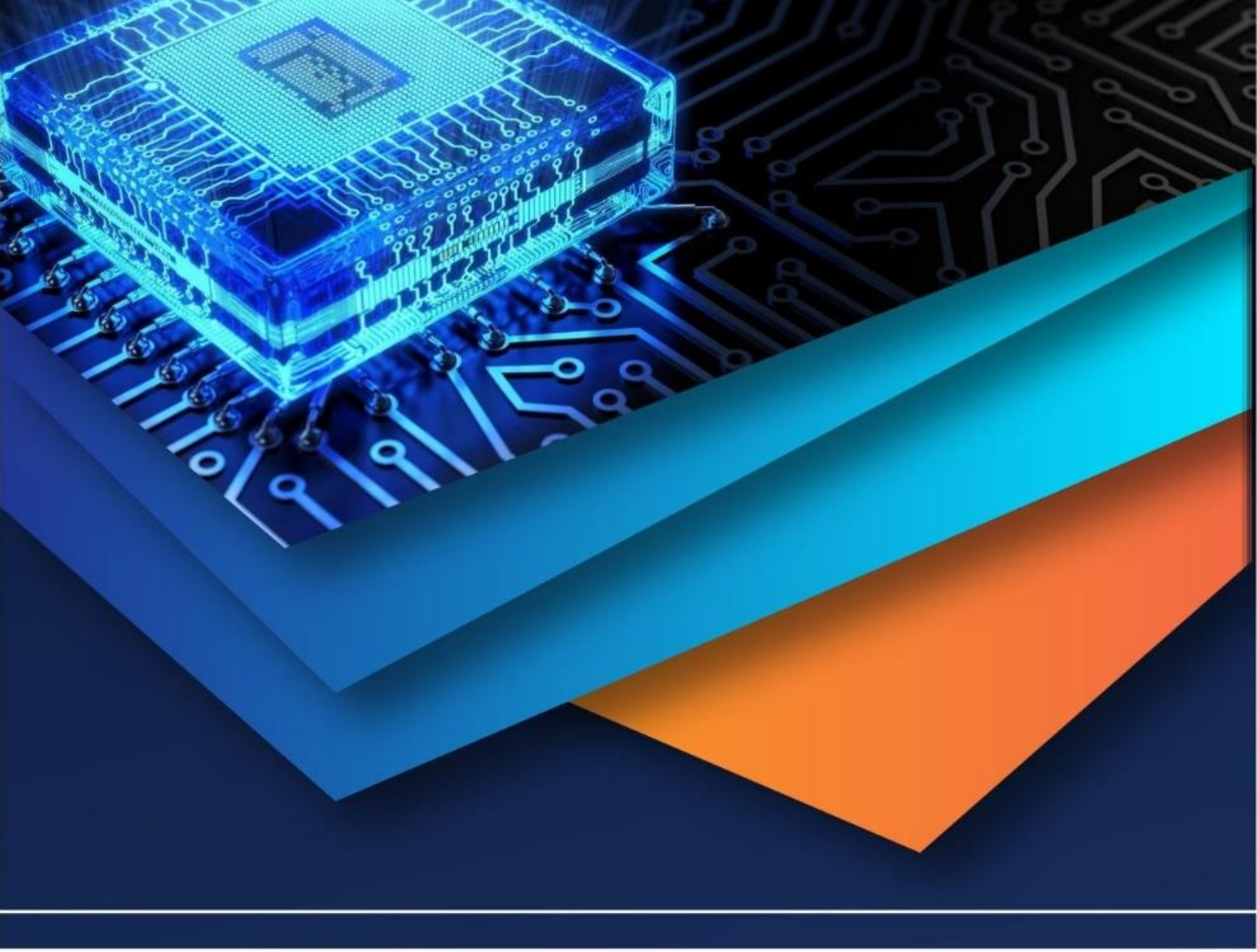
In [5], the methods used are based on CNN and LSTM which lacks in naturalness in the generated captions but are highly accurate as it uses supervised learning whereas GAN architecture has less accuracy as it follows unsupervised learning.
In [11], On Comparing CSF and the baseline, the CSF image description and the baseline model which is named as 'SAT'. On the basis of resultant captions, the objects in the image could be identified by CSF and more detailed descriptions are induced by the object-aware semantics.
By comparing two methods in [15], the RNNs word embedding dimension and hidden state dimension are set as 256 and for the multimodal method, and 0.0001 is the learning rate of multimodal method.

IV. CONCLUSION

Image captioning system is thus the most effective system which can be used for many purposes. It aids in automatic caption generation for an image in a single sentence. It can also be used to help the visually impaired people who depend on text or description to describe an image. It not only helps the blind people but also the people who cannot see the far objects. In the system which will be created CNN will be utilized for object detection and LSTM will be utilized for caption generation.

REFERENCES

[1] Sreela S R & Sumam Mary Idicula, "AIDGenS: An Automatic Image Description System using Residual Neural Network", International Conference on Data Science and Engineering (ICDSE), Pp.1-5, 2018
[2] Parth Shah, Vishvajit Bakrola & Supriya Pati, "Image Captioning using Deep Neural Architectures", International Conference on Innovations in information Embedded and Communication Systems (ICIIECS), Pp.1-4, 2017
[3] Chetan Amritkar & Vaishali Jabade, "Image Caption Generation using Deep Learning Technique", Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), Pp.1-4, 2018
[4] 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", 5th International Conference on Advanced Computing & Communication Systems (ICACCS), Pp.1-3, 2019
[5] Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, Hamid R & Arabnia, "Image Captioning with Generative Adversarial Network", 9 International Conference on Computational Science and Computational Intelligence (CSCI), Pp.1-4, 2019
[6] Smriti Sehgal, Jyoti Sharma & Natasha Chaudhary, "Generating Image Captions based on Deep Learning and Natural language Processing", 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) Amity University, Pp.1-5, 2020
[7] Pranay Mathur, Aman Gill, Aayush Yadav, Anurag Mishra & Nand Kumar Bansode, "Camera2Caption: A Real-Time Image Caption Generator", International Conference on Computational Intelligence in Data Science (ICCIDS), Pp.1-6, 2017
[8] 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" 5th International Conference on Advanced Computing & Communication Systems (ICACCS),Pp. 1-3, 2019
[9] Vaibhav Pandit, Rishab Gulati, Chaitanya Singla, Dr.Sandeep & Kr.Singh, "DeepCap: A Deep Learning Model to Caption Black and White Images", 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Pp. 1-5, 2020
[10] Jie Wu & Haifeng Hu, "Cascade recurrent neural network for image caption generation", National Natural Science Foundation of China & Science and Technology Program of Guangzhou, China, Pp.1-2, 2017
[11] Shiwei Wang, Long Lan, Xiang Zhang, Guohua Dong & Zhigang Luo, "Cascade Semantic Fusion for Image Captioning", National Natural Science Foundation of China, Pp.1-9, 2016
[12] Xiaolin Hu, Binheng Song, Jian Yang & Jianwei Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding", IEEE Transactions on Image Processing, Pp.12-15, 2018.
[13] Daouda Sow, Zengchang Qin & Mouhamed Niasse, "A sequential guiding network with attention for image captioning", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Pp.12-15, 2019.
[14] Adela Puscasiu, Alexandra Faunca, Dan-Ioan Gota & Honoriu Valean, "Automated image captioning", IEEE International Conference on Automation, Quality and Testing, Robotics (AQTR), Pp.12-15, 2020.
[15] Xiaoqiang Lu, Binqiang Wang & Xiangtao Heng, "Exploring Models and for Remote Sensing Image Caption Generation", IEEE Transaction on geoscience and remote sensing, Pp.12-15, 2017.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY