iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** VII **Month of publication:** July 2022

DOI: https://doi.org/10.22214/ijraset.2022.45988

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: 7.538
Volume 10 Issue VII July 2022- Available at www.ijraset.com

Image Description Generator using Deep Learning

Deepak R Ksheerasagar[1], Nanda Kumar T K[2], Manoj H Y[3], Shashikanth M S[4], Asst. Prof. K. N. Prashanth Kumar[5]
[1, 2, 3, 4, 5]*Department of Computer Science & Engineering, Bangalore Institute of Technology*, Bengaluru, India

Abstract: ***To recognise the context of an image and describe it in a natural language like English, the fundamental task of creating image captions uses computer vision and natural language processing techniques. To create a natural language description from an input image, image caption generation is used. Convolutional Neural Network (CNN) model and Long Short-Term Memory (LSTM) model are the two parts of this Python project that are used to implement it. The CNN-LSTM architecture combines a Convolutional Neural Network (CNN), which creates features that describe the images, with a Long Short-Term Memory (LSTM), a type of Recurrent Neural Network (RNN), which precisely structures meaningful sentences out of the generated data. The ability to automatically describe an image's content has a variety of uses, including helping visually impaired people better understand the content of images and providing more precise and condensed image information for social media.***

Keywords: CNN, RNN, LSTM, BLEU score, GPU, ResNet50

I. INTRODUCTION

A well-known topic in the field of deep learning is image caption generator, which focuses on describing an image based on the objects it contains. Combining image-scene comprehension, feature extraction, and translation of the visual representations into natural language make up this task. Understanding the language used to describe the captions is necessary for producing well-structured, logically sound descriptions.

The main challenge in this task is describing the image because the model must accurately describe how objects relate to one another as well as search for and identify the images that are present in the image.

After that, the image is described in English, necessitating not only visual comprehension but also a language model. A CNN and RNN combined network are used in our model.

RNNs are used for caption generation, while CNNs are used for image processing. LSTMs, which are particularly useful for generating the right word order, are used in RNNs.

There are many uses for image captioning, including security systems, robotic vision, and assisting the blind. Looking ahead to automated image captioning's potential applications, it is obvious that this area is both complex and highly promising. But a significant concern is achieving high accuracy for caption generation models.

The model's generated captions must identify the objects and explain how they relate to one another in the image.

The described model is built on a CNN for image classification, followed by an RNN for word generation. English captions are generated for the provided input image.

Machine learning, image processing, and computer vision have become very important and cost-effective requirements in a variety of fields and applications. These cover everything from face and iris recognition in forensics to signature recognition for authorization. Additionally, their combination is widely used worldwide in military applications. Each of these applications has particular prerequisites that may make it different from the others. Any stakeholder in such systems or models is concerned and demands that their system be quicker, more accurate than competitors, as well as less expensive and outfitted with more powerful computational capabilities. Since most of these systems are used for mission-critical purposes and the likelihood of a mistake should be extremely low, all the desired characteristics from the systems are desirable.

A computer vision-based model that is impartial and devoid of any prejudice towards anything or anyone is needed to generate a caption describing the images given to it as input.

Such systems are required to handle the complexity of problems in the modern world, such as intelligent crimes, smart city needs like smart traffic control systems, disaster control and management systems, etc.

In order to reduce the possibility of errors in such crucial tasks, such a description should be used to automate existing systems like traffic control, flood control, or surveillance systems.

In addition, surveillance could be carried out continuously without the need for human interaction.

©IJRASET: All Rights are Reserved

II. APPLICABILITY

There are many uses for image captioning in a variety of industries, including biomedicine, business, online search, and the military, among others. Social media platforms like Facebook, Instagram, and others can automatically create captions from images. We show how to use recurrent neural networks (RNNs) effectively to produce captions for the system's input images in natural language (English).

III. LITERATURE SURVEY

This paper [1] develops and assesses two techniques for automatically mapping images to descriptions in natural language. The first retrieves and transfers complete captions from database images to a query image using global image feature representations (Ordonez et al. 2011). The second extracts textual phrases from numerous visually comparable database images, providing the phrases as the cornerstones from which to build original and subject-specific captions for a query image. The main flaw in this study is that although the generated outputs are typically grammatically sound and fluid, limiting image descriptions to preexisting sentences makes it difficult for the model to adapt to novel object combinations or scenes.

The authors of this paper [2] introduce dense captioning, which differs from object detection in that it replaces the fixed number of object categories with a much larger set of visual concepts that are expressed as phrases. Dense captioning is similar to object detection in that it also requires localizing the regions of interest in an image. Based on the fully convolutional neural network-generated image feature maps, they employ a two-stage neural network to detect objects. The network first creates region proposals that are very likely to be the regions of interest using an RPN, and then uses Region-Of-Interest (ROI) pooling layers to create fixed-length feature vectors for each region proposal. The feature vectors are fed into a different network in the second stage in order to predict object categories and bounding box offsets. Exact joint training cannot be used for faster R-CNN because the gradients cannot be propagated through the proposal coordinates. Instead, it can be trained by either approximate joint training, which updates the parameters with gradients from the two components jointly, or by alternatively updating parameters with gradients from the RPN and the final prediction network. At the end of each time step, the LSTM predicts a word and uses that prediction to predict the following word. Convolutional layers are created using the VGG-16 structure, which results in feature maps that are 1616 smaller than the input image. Pretrained weights from the ImageNet Classification challenge are used, which is faster R-CNN. Only the region feature is fed into the LSTM at the first-time step, then a unique token for the beginning of a sentence, and finally, one by one, the embedded feature vectors of the predicted words. The main flaw in this paper is the excessive number of redundant computations, or the inefficient forwarding of each region for a given image separately.

In this paper [3], the authors created a deep hierarchical framework for image captioning that receives image representations with various CNN depths and divides the language model into multi-layer LSTMs. This gives the encoder-decoder a mechanism to increase its vertical depth. With this method, the multi-level semantics of language and vision can be combined for caption generation, significantly enhancing the proposed network's capacity for representation. They employ two different types of CNN image encoders: the Inception-Resnet-v2 model and a 16-layer VggNet with 13 convolutional layers. In our experiments, the used VggNet model was pre-trained using the 1.2M image ILSVRC-2012 classification training subset of the ImageNet. A new group of convolution layers, conv6, is added behind the fifth group of convolution layers, conv5, specifically to obtain different depth CNN image features. The structure of these convolution layers is the same as other convolution layers of the VggNet. Conv6's output is processed by a max-pooling layer before being connected to three additional layers that are the same kind of fully connected layers as VggNet. Finally, the sentence is produced using one of two possible methods. One is selecting the word with the highest probability as the current output at each time step, then feeding that word as an input into the subsequent calculation until the EOS word appears or the maximum sentence length is reached. The other is beam search, which searches the top-k best fractional sentences at each time step and considers them potential candidates to produce new top-k best sentences at the following time step. The process has a vanishing gradient issue because image information is only fed at the beginning of the process, which is the main disadvantage.

It [4] proposed an all-encompassing model that fully utilizes sophisticated visual and semantic data to produce descriptions. By establishing attention mechanisms, it can effectively extract features of high quality. However, it can also fully utilize the ordered memory module to take advantage of the grammar's order information. They investigated the residual attention mechanism and ordered memory module fusion network and proposed a novel attention mechanism and ordered memory module fusion network (ON-AFN), which is better able to capture visual information from images and more accurately describe their semantic content. They used the currently popular Residual Attention Network during the encoding phase, adopting a combined coarse-and-fine strategy. combining several attention modules that each concentrate on a different piece of information, allowing the model to suppress the information gradient without placing an undue strain on the training process.

The ON-LSTM ordered memory module, which can fully utilize the information about the order of neurons, enhance the expression of features in the image, and generate natural language descriptions with lower granularity and richer level, was used in the decoding stage. It uses a residual attention mechanism, which increases the model's weight parameters and lengthens training time.

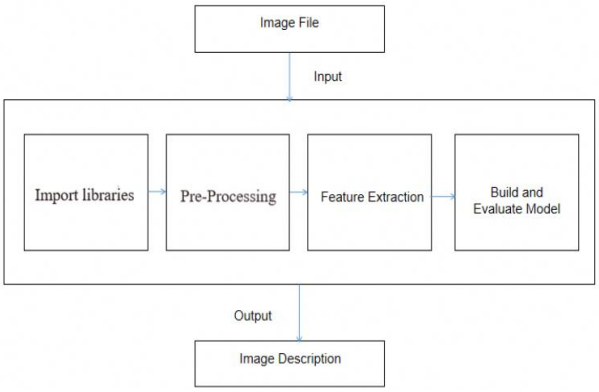
IV. PROPOSED METHDOLOGY

In order to complete this task, a CNN and an RNN are combined in the neural network architecture. The diagram for this is as follows: A sequence generator architecture like RNN or LSTM can start by converting an image into a feature vector of fixed length, which can then be used to generate a series of words or captions for the image. ResNet50 is the encoder that we have employed for the benefit of this project. The ImageNet Dataset's million images were divided into a thousand categories using a pretrained model. Since its weights are tuned to identify a lot of things that commonly occur in nature, we can use this net effectively by removing the top layer of 1000 neurons (meant for ImageNet classification) and instead adding a linear layer with the number of neurons same as number of neurons that you're LSTM is going to output. The RNN consists of a series of LSTM (Long Short-Term Memory) Cells which are used to recursively generate captions given an input image. These cells utilize the concept of recurrence and gates in order to remember information in past time steps. You can watch this or read this to understand more about the same. Eventually, the output from both encoder and decoder is merged and passed to a Dense Layer and finally an output layer which predicts the next word given our image and current sequence.

The proposed system is as follows:

1) Our first option is the graphical user interface (GUI). The user engages with the system at this point.
2) If a user is a first-time visitor, he or she must login or register.
3) After completing this, the user will have the choice to upload an image and receive a description of it.
4) After the user inputs the link or provides the text, we'll use CNN to extract features from the image and turn it into a feature vector with a fixed length.
5) Following the extraction, we pre-process the images by modifying their size, orientation, colour, brightness, and perspective. We also remove a lot of noise from the caption, such as punctuation, during this process.
6) The feature vector would then be fed to the RNN, which would then recursively generate a caption for the provided image.
7) The users would be shown the description produced by the model as the last step.

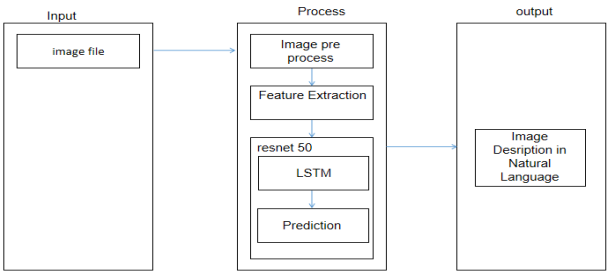
V. MODULE DECOMPOSITION



1) Import libraries: Libraries are imported in this step, including NumPy, pandas, ReNet50, Tokenizer, and seaborn. The text file's content is transformed into a panda's data frame. The dataset is in the.txt format.
2) Pre-processing and feature extraction: There are four main steps. The captions are vectorized using Kera's' Tokenizer class. Deciding on the right caption length and vocabulary size Words are converted into integer indices. Resize the photos to 224x224 and add three color channels.
3) Model Building: Model construction entails four main steps. Define the encoder network (ResNet50). The decoder network is made up of a number of LSTM (Long Short-Term Memory) Cells that are used to generate captions iteratively from an input image.

4) *Model Evaluation:* Bilingual Evaluation Understudy is also known as BLEU. It is an algorithm that compares the output quality of textual problems produced by machines and humans in an imprecise manner. On a broad scale, it compares the n-grams of a caption that was created by a human and those of a caption that was created by a machine, looking for commonalities. The better the caption quality, the more common elements there are between the corresponding n-grams of the two captions.

VI. BLOCK DIAGRAM



VII. CONCLUSION

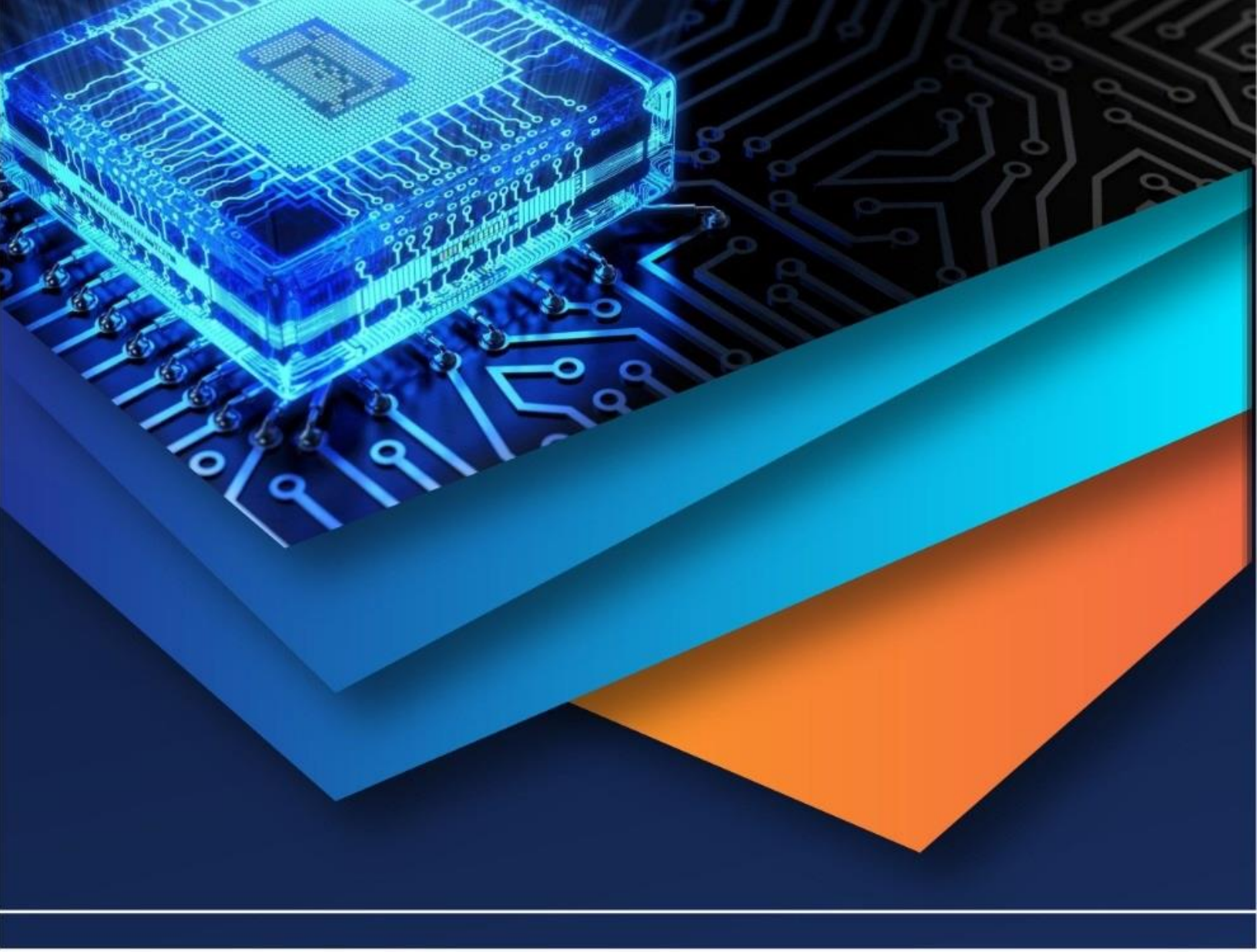
This report presents a deep learning model that successfully implements an image caption generator tool using Long-Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs). The model-generated image descriptions focused on various objects within the images to provide detailed descriptions of the images. Additionally, it is made sure that the model created for caption generation does not contain any biases or erroneous assumptions based on any of the features present in the image. Additionally, it was observed that the trained model could identify the orientations of motions made by different objects in images. The features from the input images are successfully extracted by the presented model, and they are also given a natural language description.

This report provides a thorough analysis of recent advances in the fields of artificial intelligence, deep learning, computer vision, natural language processing, and image processing. They improved our comprehension of and adoption of various feature extraction techniques as well as model evaluation techniques. The presented model is evaluated using the BLEU score, and a suitable accuracy value was obtained for the model. As the application domains of this model require critical use and scalability to fulfil its purpose, the state-of-the-art works also highlighted the need for improvement in the techniques used to automatically describe an image. When building a deep learning model, tools and packages like Keras, NumPy, the Nltk package, etc., offer a variety of functions to guarantee a smooth training and testing process. New mages are also described by the model with a level of accuracy that is acceptable after training. It is thus a practical and scalable model for tasks involving image generation.

REFERENCES

[1] Vicente Ordonez1, Xufeng Han1, Polina Kuznetsova2, Girish Kulkarni2, “Large Scale Retrieval and Generation of Image Descriptions” in Springer Science+Business Media New York
[2] Linjie Yang, Kevin Tang, Jianchao Yang, Li-Jia Li, “Dense Captioning with Joint Inference and Visual Context” in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
[3] Xinyu Xiao, Lingfeng Wang, Kun Ding, Shiming Xiang, and Chunhong Pan, “Deep Hierarchical Encoder-Decoder Network for Image Captioning” in Journal of IEEE Transactions on Multimedia volume 21, 2018.
[4] Akshansh Chahal, Manshul Belani, Akashdeep Bansal, Neha Jadhav, and Meenakshi Balakrishnan, “Template Based Approach for Augmenting Image Descriptions” in International Conference on Computers Helping People with Special Needs, 2018
[5] Jun Yu, Member, Jing Li, Zhou Yu, Qingming Huang, “Multimodal Transformer with Multi-View Visual Representation for Image Captioning”, in IEEE Transactions on Circuits and Systems for Video Technology, 2019
[6] Zhenyu Yang, Qiao Liu and Guojing Liu, “Better Understanding: Stylized Image Captioning with Style Attention and Adversarial Training” in Computer and Engineering Science and Symmetry/Asymmetry, 2020.
[7] Ying Hua Tan, Chee Seng Chanm, “Phrase-based image caption generator with hierarchical LSTM network”, in Neurocomputing, Volume 333, 14 March 2019, Pages 86-100.
[8] Md Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga, “Bi-SAN-CAP: Bi-Directional Self-Attention for Image Captioning” in Conference on Digital Image Computing: Techniques and Applications (DICTA), 2020
[9] Chenchen Lu, Zhaowei Qu, Xiaoru Wang, Heng Zhang, “ON-AFN: Generating Image Caption based on the Fusion of Residual Attention and Ordered Memory Module” in IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), 2020.
[10] Chunlie Wu, Shaozu Yuan, Haiwen Cao, Yiwei Wei, AND Leiquan Wang, “Hierarchical Attention-Based Fusion for Image Caption With Multi-Grained Rewards” in IEEE Journal, volume 8, page 57943-57951, 2020.

[11] 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 Proc. CVPR, Jun. 2018, pp. 6077–6086.
[12] S. Banerjee and A. Lavie, ''Meteor: An automatic metric for MT evaluation with improved correlation with human judgments,'' in Proc. ACL Work-shop Intrinsic Extrinsic Eval. Measures Mach. Transl., 2005, pp. 65–72.
[13] C. Wu, Y. Wei, X. Chu, S. Weichen, F. Su, and L. Wang, ''Hierarchical attention-based multimodal fusion for video captioning,'' Neurocomputing, vol. 315, pp. 362–370, Nov. 2018.
[14] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio, An Actor-Critic Algorithm for Sequence Prediction. Cambridge, MA, USA: MIT Press, 2016.
[15] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig, ''From captions to visual concepts and back,'' in Proc. CVPR, Jun. 2015, pp. 1473–1482.
[16] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, ''Every picture tells a story: Generating sentences from images,'' in Proc. ECCV, 2010, pp. 15–29.
[17] S. Hochreiter and J. Schmidhuber, ''Long short-term memory,'' Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997.
[18] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, ''Long-term recurrent convolutional networks for visual recognition and description,'' in Proc. CVPR, Jun. 2015, pp. 2625–2634.
[19] J. Lu, C. Xiong, D. Parikh, and R. Socher, ''Knowing when to look: Adaptive attention via a visual sentinel for image captioning,'' in Proc. CVPR, Jun. 2017, pp. 375–383.
[20] Y. J. J. Yang Wang Mao, W. Xu, and A. Yuille, ''Deep captioning with multimodal recurrent neural networks (M-RNN),'' in Proc. ICLR, 2015.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)