iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH
IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 Issue: XI Month of publication: November 2025

DOI: https://doi.org/10.22214/ijraset.2025.75144

www.ijraset.com

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

Image Caption Generator: A Comprehensive Review

Himanshi Yadav[1], Khushi Rastogi[2], Dr. Rajan Prasad[3], Er. Anurag Chauhan[4]
[1, 2]*Student, Dept. of Computer Science & Engineering, Shri Ramswaroop Memorial College of Engineering & Management, Lucknow, India*
[3, 4]*Assistant Professor, Dept. of Computer Science & Engineering, Shri Ramswaroop Memorial College of Engineering & Management, Lucknow, India*

Abstract: In recent years Computer vision has drastically advanced in the field of image processing. Image captioning, which involves automatically generating one or more captions to comprehend an image's visual information, has benefited from advancement in image detection. In this paper we aim to prove that combination of existing methods can efficiently improve the performance in image detection. The approach involves generating meaningful captions by combining computer vision and neural language processing. The technologies used are Convolutional Neural Networks (CNNs) to extract image features and Long Short-Term Memory (LSTM) networks with attention mechanisms to produce coherent sentences. A trained model that has been trained with algorithm and Flickr8K, Flickr30K dataset will produce the caption. We talk about Python, TensorFlow and deep learning frameworks for making this paper. The main use case of this research is to help visually impaired to understand surrounding environment. It can be used in hospitals to treat patients with any neurological conditions. The paper reviews the previously done researches and enhance the model already present and discuss their advantages, disadvantages and future scopes.

Keywords: Image Captioning, CNN, LSTM, NLP, Deep Learning.

I. INTRODUCTION

In our daily lives, we come across countless images on social media, websites, newspapers, and advertisements. As humans, we can instantly understand what these pictures show, but computers cannot do this naturally. To help machines describe images on their own, a technology called Image Caption Generation is used. It enables a computer to look at an image and create a sentence that explains it —: for example. A girl riding a bicycle or A dog sleeping on a bed. Image Captioning is a fascinating part of Artificial Intelligence (AI) that brings together two key areas — Computer Vision and Natural Language Processing (NLP). Computer Vision helps the machine evaluate and recognize objects or scenes in an image. NLP helps it form proper and meaningful sentences in human language. Creating accurate captions for images is a difficult task because it requires understanding both visuals and language. To improve caption quality, researchers use Deep Learning methods. These include Convolutional Neural Networks (CNNs), which is used in identifying objects within an image, and Long Short-Term Memory (LSTM) as Recurrent Neural Networks (RNNs), which help in generating text descriptions word by word. Some systems also use VGG16, a well-known CNN model, to recognize detailed visual features. By combining these models, the system can first extract features from an image and then generate a caption that correctly represents it. For training, datasets like Flickr8k are often used. This dataset contains thousands of images, each paired with several human-written captions. The model learns how images and words relate, so it can later produce captions for new images that it has never seen before. Image caption generators have applications in accessibility for visually impaired user, search engine Optimization (SEO), Content Organization, and social media, creating descriptive text for images. They are used in areas like digital libraries, and educational platforms to make visual information more understandable and searchable.

II. LITERATURE SURVEY

It is difficult to generate textual captions from any given image using artificial intelligence. It requires two methods first understand the content of the image using computer vision and second turn the understanding of image into right words using large language processing.

We achieve this by using techniques such as Long -Short -Term Memory, Convolutional Neural Networks (CNN), and images and their descriptions which is understood by humans. We found that our model produced results by getting trained on Flicker dataset.

There are many researches done with videos and subtitles but an overview should be given before starting

1) CANONICAL CNN-LSTM ENCODER AND DECODER: The canonical pipeline was widely used after early works that trained an LSTM language model conditioned on CNN image features. In this process, the final CNN feature vector is used to initialize the LSTM hidden state or fed as an input token; then the LSTM generates the caption token by token by increasing the similarities of captions during training. This simple architecture established a strong baseline and clarified how to combine vision and language in a trainable model.[1]
2) ALIGNING WORDS TO IMAGES (ATTENTION MECHANISMS): A major improvement in image captioning was the use of attention mechanisms in the decoder. Attention allows the model to focus on important regions of an image while generating each word, leading to more accurate and detailed captions.[2]
3) TRAINING OBJECTIVE AND IMPROVEMENTS: It uses maximum likelihood to optimize prediction of next word. Reinforcement-learnings were introduced to directly optimize non-differentiable sequence metrics: notable is Self-Critical Sequence Training (SCST), which uses the model's own greedy output as a baseline for policy-gradient updates. SCST and similar approaches reduced exposure bias and raised scores on task metrics.[3]
4) DATASETS AND EVALUATION: Datasets Flick8k, MS COCO are used because captioning relies on automatic metrics. Each has strength and limitations. Using several dataset together can give a balanced performance.[4]
5) BEYOND CNN AND LSTM: CNN–LSTM are important frameworks (simple, interpretable, good for learning the basics), the field has shifted toward fully-attentive models that replace recurrent decoders with self-attention, and sometimes replace CNN encoders with Vision Transformers or region features. Transformer variants generally achieve superior results by modeling relationships among images and allowing more flexible cross-modal attention. These newer models often outperform classic CNN–LSTM systems in standard benchmark, though they are more resource similar.
6) LEARNING VS DEEP LEARNING: The training data comes with label in supervised learning. Many image captioning methods use reinforcement learning and GAN based approach. [7]
7) COMPOSITIONAL ARCHITECTURE VS ENCODER -DECODER ARCHITECTURE:_Some methods use vanilla encoder and decoder to generate captions while others use multiple networks. [8]

III. RESEARCH OBJECTIVES

A. *General Objective*

The main objective of this research is to create a system that can generate meaningful caption for any given image automatically in human-like language. It uses a combination of computer vision and natural language processing to understand what is in the image and form meaningful sentences. The goal is to make captions accurate, grammatically correct and relevant.

B. *Specific Objectives*

1) Analyzing existing image captioning model and the most effective architectures (like encoder-decoder, attention mechanisms).
2) CNN will extract the key features from images and represent them in format which is suitable for generating captions.
3) Design and implement a model capable of generating descriptive captions for wide range of images
4) Evaluating the model's performance using standard metrices such as BLEU.
5) To enhance caption quality by integrating attention mechanisms or pre-trained language models.

IV. PROPOSED METHODOLOGY

Model: Encoding – Attention– Decoding
Encoder (Vision): Pretrained CNN backbone to extract spatial feature maps (e.g., 7×7×2048).
Attention module: Spatial attention or cross-attention (Transformer-style) to focus on relevant image regions for each word.
Decoder (Language): Either LSTM with attention or a Transformer decoder that generates tokens autoregressively.
Training: Start with cross-entropy loss (teacher forcing), add scheduled sampling to reduce exposure bias, then optional or BLEU-rewarded (e.g., Self-Critical Sequence Training).
The datasets are trained in order to obtain accuracy in future
in case new data is fed to the model.
Preprocessing: Resize images, normalize, tokenize captions, build vocabulary, convert captions to integer sequences, pad/truncate.
Evaluation: Automatic metrics — BLEU, METEOR, plus small-scale human evaluation for fluency & relevance.

A. Workflow Steps

1) Dataset collection and preparation: Use of Flick8k dataset containing clean captions, lowercase, removing punctuation. Building vocabulary.
2) Preprocessing of image: Resize of image with ImageNet. Augmenting only during training when needed.
3) Extraction of Features (Encoding): Passing image through pretrained CNN and obtaining spatial feature map or global vector.
4) Preprocessing of text: Tokenization of caption, converting to indices of fixed length.
5) Attention and Decoding: Compute attention weights on image features in each decoding step and combine to form context vector.
6) Training and loss: Using teacher forcing with teacher friendly schedule. After completing baseline training, apply RL fine training.
7) Evaluating: Generating captions on validation, computing BLEU to perform human evaluation on random subset.
8) Deploy: Exporting the module and serve via Fast API with an endpoint that accepts images and returns captions.

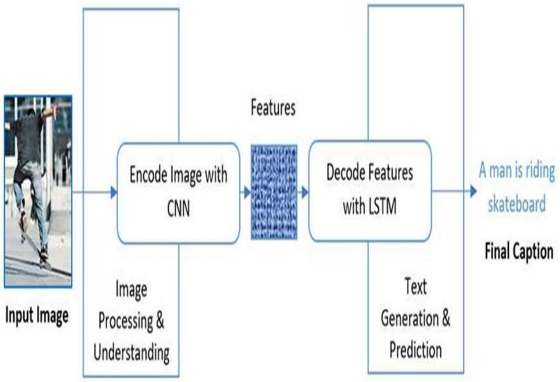

Input Image
Encode Image with CNN
Image Processing & Understanding
Features
Decode Features with LSTM
Text Generation & Prediction
A man is riding skateboard
Final Caption

Fig.1 Block view of proposed methodology

A. Model Overview

The model is trained to get the probability. It is designed to use a dictionary created by teaching information. The input image is fed into a deep neural network (CNN), which facilitates the detection of objects in the image. The language is created as shown in Figure. Recurrent Neural Networks (RNN) take image encodings and use them to create image-related sentences.[9] This model can be compared to the translation RNN model.

B. Convolutional Neural Network

It is a Deep Learning algorithm which can take in an input image, assign important weights and biases to various objects in the image and be able to differentiate one from the other. The pre-processing required in a Conv-Net is much lower as compared to other classification algorithms. Convolutional Neural networks are distinguished deep neural networks which can process the data that has input shape like a 2D matrix. Images are easily shown as a 2D matrix and CNN is very useful in working with images. It scans images from left to right and top to bottom to pull out important features from the image and combines it to classify images.

C. Long- Short -Term Memory

LSTM (Long Short-Term Memory) is an advanced version of Recurrent Neural Networks (RNNs) developed to effectively learn patterns from sequential data while collecting important long-term information. It works on the limitation of traditional RNNs, which often forget earlier inputs due to the vanishing gradient problem. LSTM networks achieve this through a unique internal structure that includes three gates — the forget gate, input gate, and output gate — which collectively decide what information should be remembered, updated, or removed at each time step. This mechanism allows the model to maintain context over longer sequences, making it suitable for applications such as language modeling, speech recognition, and image caption generation. In an image captioning system, for example, an LSTM acts as the decoding agent that generates detailed description of text based on visual features extracted from an image, enabling the model to produce meaningful captions that connect visual understanding with language generation.

D. VGG16

Simonyan, K. & Zisserman, A. (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition," arXiv:1409.1556, and other standard deep learning resources.[16] VGG-16 is a deep convolutional neural network (CNN) architecture developed by the Visual Geometry Group (VGG) at the University of Oxford. It is known for its simplicity and using small 3×3 convolutional filters throughout the network while increasing the depth to 16 weight layers — hence the name VGG-16. The model follows a straightforward pattern: multiple convolutional layers are stacked together, a max-pooling layer follows it to reduce spatial dimensions, and in the end connected to three fully connected layers for classification. Even though it has simple design, VGG-16 has achieved outstanding performance on large-scale image recognition tasks such as ImageNet. The use of filters allows the network to capture fine details of minute image while keeping computational efficiency.Due to its strong feature extraction capability, VGG-16 is widely used in computer vision tasks like face recognition, and image captioning. In image caption generation, for instance, VGG-16 is commonly used as the encoder that extracts meaningful visual features from an image, which are then passed to a decoder such as an LSTM to generate meaningful captions.

"progress". It is simple to implement in functions and loops

E. Encoder Model

This model is basically responsible for processing the captions of images that are given to it during training. The output is in vector of size 1*256 that would be an input to the decoder.

The important part of this model is LSTM layer. This layer allows model to learn to generate valid sequences or generate words with highest possibility of occurance.ReLU is the activation function used.

To compare VGG+LSTM and VGG+GRU a particular layer will be replaced by GRU which is Gated Recurrent Units.[GRU has similar output space, the only difference will be in the encoder part.

F. Decoder Model

This model is used to combine encoder model and feature extraction model and produce the required output ehich is predicted word given as image and sentence is generated till that time.

The dimension is 256 and the output of the model is passed inside a dense layer which will use activation function(ReLU). One more dense layer is added and the size of vocabulary will be similar to output space. The Flick 8k dataset had vocabulary size of 7579 and the activation function was softmax. The output of the decoded layer is predicted word.

The input is image and output is predicted caption. After the generation of caption BLEU score is calculated for each output.

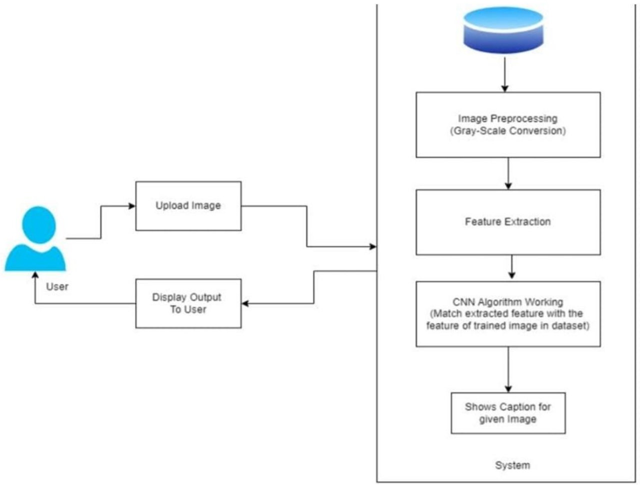


Fig 2. System Architecture

G. Libraries

1) *Keras:* To feed the Keras module original input are used named as Image Data Generator, it will change and modify the input and return a newly formed transformed data.
2) *Tensorflow:* It has a special ability to identify the picture and the images that are recognized are kept in a separate folder. This algorithm good for security. It has simple implementation with similar images. The images which are relevant are stored in dataset image and should be loaded.

3) *Pillow*: It has all the core processing features of images. Images can be rotated, modified and resized. Histogram function can be used to extract statistical data from images which can be used to perform enhancement and analysis.
4) *Numpy:* It is used to perform many image analysis operations as arrays are used to represent images.
5) *Tqdm:* It creates progress meters and progress bars. It is an abbreviation of Arabic word “taqqadum” meaning

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have discussed about the design and implementation of an Image Caption Generator model using Long Short-Term Memory (LSTM) and Convolution Neural network (CNN)networks. The CNN was used to extract visual features from the input images, while the LSTM generated grammatically correction based on those feature.

The model successfully bridge the gap between natural language processing and computer vision by converting visual information into meaningful sentences. The involvement of the attention mechanism reputed the model’s ability to focus on image regions, improving caption accuracy. In the future, the model can further improve by training on larger dataset to increase caption accuracy and contextual understanding. Caption generation can make the system accessible to the audience.

The use of CNN architectures such as EfficientNet or Vision Transformers, along with Transformer-based language models like BERT or GPT, can further increase caption quality and efficiency. The model can be integrated with IoT devices for the visually impaired, and real-time surveillance systems, making it a valuable tools for accessible and smart automation. Implementing real-time image captioning on devices would also make the system more efficient and practical for real-world application.

VI. COMPARATIVE ANALYSIS

S.No	Author(s) & Year	Objective / Research Focus	Methods / Techniques Used	Dataset(s) / Experimental Setup	Key Findings / Results	Limitations / Gaps Identified	Future Scope / Remarks
1	D. Zhu, J. Chen, K et al. 2023	To introduce a image caption generator model	LSTM+CNN Encoder& Decoder model	Ms COCO, Flickr8k datasets	Generated grammatically captions from image	Limited vocabulary	Improve contextual understanding and fluency
2	T. Nguyen, R. Marten et al. 2023	To combine attention mechanism in image caption	Attention-based LSTM+CNN model	Ms COCO dataset	Improved caption accuracy by focusing on image feature	High computational	Creates lightweight attention architectures
3	D. Simig, S. Ganguli et al. 2023	To improve attention with top down and bottom up	Region-based LSTM+CNN framework	Ms COCO datasets	Improve image-region and caption detail	Requires complex region	Proposal generation process
4	J. Donahue et al. 2022	To improve contextual representation in image caption	Transformer-based Memory Network	Ms COCO, Flickr30k datasets	High-quality captions	High model complexity	Optimize transformer models for efficiency and accuracy
5	Sharma et al., 2021	To improve caption generation using reinforcement learning	RL-based LSTM+CNN fine-tuning	Flickr30k, Ms COCO datasets	Improved the fluent sentence generation	Slow merging and tuning difficulty	Combine supervised and RL-based learning
6	Li et al.2022	To combine understanding for better captions	Visual-Linguistic BERT	Ms COCO	Improved visual-text and accuracy	Requires large-scale	Build compact and efficient caption models
7	J. Li, D. Li. 2023	To generate captions without explicit retraining	CLIP + GPT-based hybrid model	Open Images, CC12M datasets	Generate captions for images	Limited accuracy and contextual	Combine CLIP language models
8	Wang et al. 2024	To develop multilingual image caption	Transformer with multilingual image caption	COCO-CN datasets	Produced multilingual captions of multiple datasets	Language towards English	Create multilingual datasets
9	W. Zhang, P. Torr et al. 2022	To design real-time captioning for accessibility	Lightweight GRU+CNN models	Real-world dataset	Generated real-time captions with low latency	Limited descriptive quality	Improve descriptive while maintaining speed
10	Kim & Lee.2025	To provide explain in caption generation	Vision Transformer + Explainable AI	COCO, Flickr30k datasets	Visualized attention maps for transparency	Tradeoff between explainability and accuracy	Develop efficient transformer frameworks

REFERENCES

[1] Hiba Ahsan, Daivat Bhatt, Kaivan Shah, and Nikita Bhalla. 2021. Multi-modal image captioning for the visually impaired. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop. 53–60.

[2] Manuele Barraco, Marcella Cornia, Silvia Cascianelli, Lorenzo Baraldi, and Rita Cucchiara. 2022. The unreasonable effectiveness of clip features for image captioning: An experimental analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4662–4670.

[3] A. Abbas, K. Tirumala, D. Simig, S. Ganguli, and A. S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. arXiv preprint arXiv:2303.09540, 2023

[4] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35:23716–23736, 2022.

[5] H. Bansal and A. Grover. Leaving reality to imagination: Robust classification via generated datasets. arXiv preprint arXiv:2302.02503, 2023.

[6] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. arXiv preprint arXiv:2304.14108, 2023.

[7] R. He, S. Sun, X. Yu, C. Xue, W. Zhang, P. Torr, S. Bai, and X. Qi. Is synthetic data from generative models ready for image recognition? arXiv preprint arXiv:2210.07574, 2022.

[8] X. Hu, Z. Gan, J. Wang, Z. Yang, Z. Liu, Y. Lu, and L. Wang. Scaling up vision-language pre-training for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 17980–17989, 2022.

[9] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In International Conference on Machine Learning, pages 4904–4916. PMLR, 2021.

[10] J. Li, D. Li, C. Xiong, and S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning, pages 12888–12900. PMLR, 2022.

[11] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson. Model dementia: Generated data makes models forget. arXiv preprint arXiv:2305.17493, 2023.

[12] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597, 2023.

[13] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In International Conference on Machine Learning, pages 23318–23340. PMLR, 2022

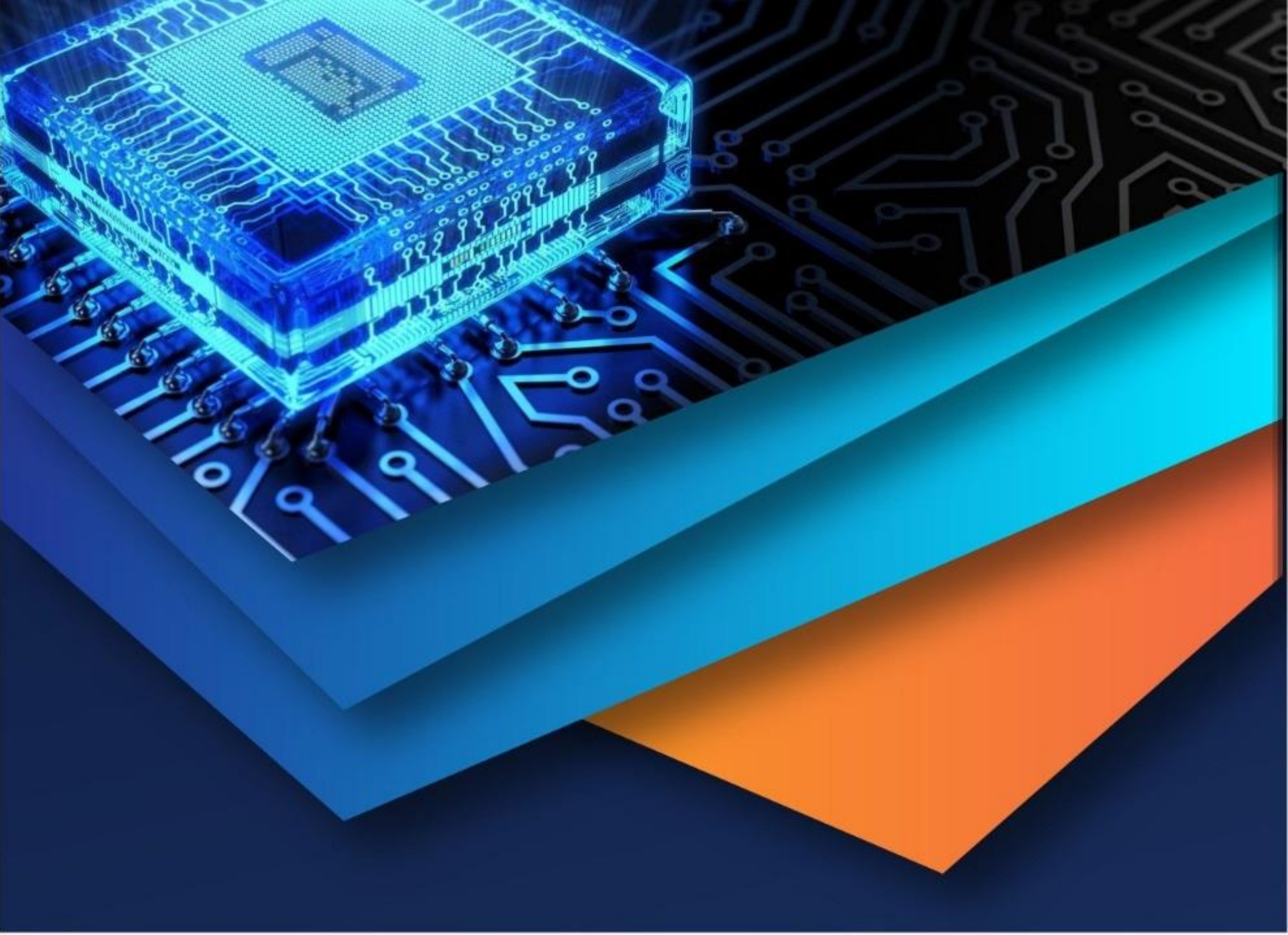
[14] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao. Simvlm: Simple visual language model pretraining with weak supervision. arXiv preprint arXiv:2108.10904, 2021.

[15] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu. Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917, 2022.

[16] Simonyan, K. & Zisserman, A. (2015). “Very Deep Convolutional Networks for Large-Scale Image Recognition,” arXiv:1409.1556, and other standard deep learning resource

[17] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny. Chatgpt asks, blip2 answers: Automatic questioning towards enriched visual descriptions. arXiv preprint arXiv:2303.06594, 2023.

[18] González-Chávez, O.; Ruiz, G.; Moctezuma, D.; Ramirez-delReal, T. Are metrics measuring what they should? An evaluation of Image Captioning task metrics. Signal Process. Image Commun. 2024, 120, 117071.



doi crossref

10.22214/IJRASET

INDEX COPERNICUS

45.98

ISRA JIF

IMPACT FACTOR: 7.129

TOGETHER WE REACH THE GOAL

IMPACT FACTOR: 7.429

ISI

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)