iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.70145

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 13 Issue V May 2025- Available at www.ijraset.com

Automated Image Caption Generator Using Deep Learning

G. Krishnaveni[1], Dr. G. Srinivasarao[2], K. Sita[3], R. Harika Priya[4], K. Usha Rani[5]
[1]Assistant Professor, [2]Professor, [3, 4, 5]U.G. Students, Department of Electronics and Communications Engineering, Bapatla Women's Engineering College, Bapatla, Andhra Pradesh, India

Abstract: ***One of the most important tasks in computer vision and natural language processing is the automatic creation of image captions.. This paper presents an approach to automatically generate descriptive captions for images by combining Convolutional Neural Networks (CNNs) and Inception V3 architecture. The proposed system utilizes a pre-trained Inception V3 model to extract high-level features from input images. These extracted features are then passed to a Recurrent Neural Network (RNN), specifically an LSTM (Long Short-Term Memory) network, to generate coherent and contextually relevant captions. Inception V3, a deep convolutional neural network designed for large-scale image classification, serves as the feature extractor. It helps capture rich spatial hierarchies within the images, making it highly effective for understanding complex visual information. The LSTM network, on the other hand, is used to model the sequence of words in the caption, ensuring grammatical correctness and semantic accuracy. The system is trained on a large dataset of images paired with human-generated captions, such as the MS-COCO dataset, to ensure robust learning. The proposed method is evaluated based on its performance in generating captions that are semantically and syntactically appropriate. The model's performance is compared to other existing image captioning methods, demonstrating its effectiveness in generating descriptive and accurate captions for unseen images. This work highlights the synergy between CNNs for visual feature extraction and LSTM networks for sequence generation, offering a promising solution for tasks requiring image-to-text conversion, including image retrieval, content-based indexing, and accessibility applications.***

Keywords: ***Recurrent Neural Network (RNN), Convolutional Neural Network (CNN), Long Short Term Memory(LSTM),Caption Generation, Image Preprocessing, Natural Language.***

I. INTRODUCTION

With the rapid advancements in computer vision and natural language processing (NLP), the task of automatically generating image captions has become one of the most important research areas. The ability to automatically describe images with natural language is essential for applications like image retrieval, accessibility tools for visually impaired users, and social media content management. Image captioning systems allow machines to see, understand, and interact with the visual environment by bridging the gap between written representation and visual input. Traditional image processing approaches generally depended on handmade features and shallow models to interpret images. However, the Convolutional Neural Networks (CNNs), in particular, have revolutionised the field of deep learning by making it possible to automatically learn hierarchical image attributes. CNNs, particularly those with deep architectures like Inception V3, are extremely capable of extracting complex, high-level features from images, which can then be used for various tasks such as classification, object detection, and, more importantly, image captioning. Inception V3, a state-of- the-art CNN architecture, is well-suited for image captioning due to its ability to capture fine-grained spatial information and contextual details from images. By leveraging this powerful feature extractor, the task of image captioning becomes more efficient and accurate. However, while CNNs excel at visual understanding, they are not designed to generate sequences of text. Recurrent neural networks (RNNs), and more especially Long Short-Term Memory (LSTM) networks, are useful in this situation .LSTMs are capable of modeling sequential data, making them ideal for generating grammatically correct and contextually meaningful captions from image features. This research proposes an end-to-end image captioning system by combining Inception V3 for feature extraction and LSTM for caption generation. The system is designed to automatically generate descriptive captions for images, offering an efficient solution for a wide range of practical applications. The model is trained using large-scale image- caption datasets, ensuring its ability to generalize across various image types and captioning styles. This combination of CNN and LSTM models not only improves the accuracy of the captions but also enables the generation of more contextually relevant and coherent descriptions. The goal of this paper is to present a robust and scalable method for automatic image captioning, leveraging the strengths of both CNNs for feature extraction and LSTMs for natural language generation. We intend to enhance image performance and applicability by employing this technique. captioning systems, offering a promising solution for practical implementations in real-world scenarios.

II. LITERATURE SURVEY

1) *["BLEU: a Method for Automatic Evaluation of Machine Translation," Proceedings of the Association for Computational Linguistics' (ACL) 40th Annual Meeting, Salim Roukos, Todd Ward, Wei-Jing Zhu, andKishore Papineni Philadelphia 1–18 pages, July 2002.*

The BLEU (Bilingual Evaluation Understudy) metric, introduced by Papineni et al., is an automatic method for evaluating machine translation (MT) quality by comparing candidate translations to human reference translations. It relies on modified n-gram precision, ensuring accurate word choice without overuse, and applies a brevity penalty to discourage excessively short translations. Unlike recall-based approaches, BLEU focuses on closeness to human translations using a geometric mean of precision scores for up to 4-grams. Extensive testing showed a high correlation (0.96–0.99) with human evaluations, making it a fast, cost-effective, and language-independent alternative to manual assessment. Though primarily effective for corpus-level evaluation, BLEU has become a standard benchmark in MT research, enabling rapid system improvements.

2) *Richard S., Kyunghyun Cho, Aaron Courville, Ryan Kiros, Michael Lei Ba, and Kelvin Xu.Salakhutdinov, Ruslan. "Show, attend, and tell*

Neural image caption generation with visual attention," by Yoshua Benjio Zemel, ICML'15Proceedings of the 32nd International Conference on Machine Learning – Volume 37, Pages 2048- 2057.Feb.2015The article "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" describes a method for captioning photographs.model that enhances the quality of generated captions through the usage of an attention mechanism.The model is based on an encoder-decoder framework where a convolutional neural network (CNN) extracts image characteristics, and descriptive text is produced by a recurrent neural network (RNN), more precisely an LSTM network.The key innovation is the use of an attention mechanism that dynamically focuses on different parts of an image while generating each word in the caption, allowing for more context-aware descriptions. The paper demonstrates that this attention-based approach significantly improves captioning performance over previous methods, achieving state-of-the- art results on benchmark datasets

3) *R.Staniute and D. Sesok, "A Comprehensive Review of the Literature on Image Captioning," Applied Sciences, vol. no. 2-20, 16 March 2019*

The paper "A Systematic Literature Review on Image Captioning"by Raimonda Staniutė and Dmitrij Šešok provides a detailed review of advancements in image captioning from 2016 to 2019. It examines common techniques, challenges, and evaluation methods, highlighting inconsistencies in result comparisons. The study categorizes various approaches, including encoder-decoder models, attention mechanisms, and the use of semantics and novel objects. Popular encoders like ResNet and VGG-16, along with LSTM-based decoders, have shown strong performance. Challenges such as dataset limitations, lack of standardization in evaluations, and difficulty in generating human-like captions are discussed. The review emphasizes the importance of comparing new models against the latest research rather than older benchmarks, offering insights to guide future developments in image captioning.

III.EXISTING METHOD

An Automated Image Caption Generator using CNN, LSTM, and ResNet is a deep learning system[1] that generates descriptive text for images by combining computer vision and natural language processing (NLP)[2]. The Convolutional Neural Network (CNN) extracts visual features from an image, while the Long Short-Term Memory (LSTM) network processes these features to generate a coherent sentence. ResNet (Residual Network), a deep CNN architecture, is used for image feature extraction, as it efficiently captures high-level visual details. This approach is widely used in applications such as automatic image tagging, AI- powered content creation, accessibility tools for the visually impaired, and image- based search engines.

The working mechanism involves first passing an image through a pre-trained ResNet-50 model, which removes the fully connected layers and extracts a 2048- dimensional feature vector. These features are then fed into an LSTM-based sequence model, which is trained to generate captions based on word embeddings and previous words in the sequence. The dataset used for training, such as the COCO (Common Objects in Context) dataset, consists of images annotated with multiple human- written captions. The model learns to associate images with their descriptions through supervised learning, where captions are converted into numerical sequences, tokenized, and padded for uniformity[6]. The final output is a natural language description of the image, generated word by word until the end token is reached.

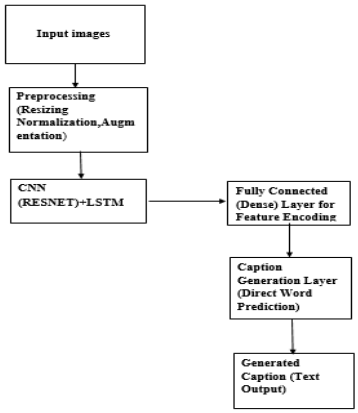


Fig 1. Block Diagram for Existing Method

Despite its effectiveness, the Automated Image Caption Generator has several drawbacks. One major limitation is its inability to fully understand the context and relationships between objects in complex scenes[7]. Since the model relies on statistical patterns rather than true comprehension, it can produce generic, inaccurate, or incomplete captions, especially for images containing abstract concepts or multiple interacting objects. Furthermore, if an image contains elements that were underrepresented in the training dataset, the model may fail to describe them correctly. This issue can lead to misleading captions or descriptions that omit crucial details. Another challenge is the high computational cost and data dependency. Training such deep learning models requires large annotated datasets and significant computational power (GPUs/TPUs)[3]. The model is also susceptible to bias, as it learns from pre- existing datasets that may not be diverse or balanced, leading to stereotypical or culturally biased captions. Additionally, hallucination errors—where the model describes objects that are not actually present—can occur due to dataset biases. To overcome these challenges, researchers are exploring transformer-based models like Vision Transformers (ViTs)[14] and multimodal AI architectures, which improve contextual understanding and reduce errors.

IV. PROPOSED METHOD

An automated image caption generator using deep learning integrates computer vision and natural language processing to generate meaningful descriptions for images. The proposed method leverages InceptionV3, a powerful convolutional neural network (CNN), for feature extraction, and an LSTM-based sequence model for text generation[4]. The MS-COCO dataset, which contains images annotated with multiple human-generated captions, is used for training and evaluation.. This approach helps the model to learn rich picture representations and translate them into natural language descriptions. Before feature extraction, images must undergo preprocessing to ensure consistency. They are sized to 299×299 pixels, transformed to RGB format, and balanced by scaling cell values to a range of [0,1]. Additionally, data augmentation techniques such as random cropping, flipping, and color jittering can be applied to improve the model's robustness and generalization. Once preprocessed, images are passed through InceptionV3, which is pre-trained on ImageNet, to extract deep visual features[8]. Instead of using the final classification layer, the fully connected layers are removed, and the 2048-dimensional feature vector from the global average pooling layer is used as a compact representation of the image. These extracted features capture crucial object-level and spatial information, which serves as input to the caption generation model. The captioning process begins with text preprocessing, where captions from the MS-COCO dataset are cleaned by removing punctuation, converting text to lowercase, and filtering out rare words to limit vocabulary size.

Captions are then tokenized, converting words into numerical indices based on a word-to-index mapping. Each caption begins with a start token ("<start>"), while the conclusion of the sequence is indicated with an end token ("<end>").To ensure uniformity in training, captions are either padded or truncated to a fixed length. Additionally, word embeddings such as BLUE, GloVe[5] or trainable embeddings are used to convert words into dense vector representations, enabling the model to understand semantic relationships between words. The LSTM-based decoder is responsible for generating captions based on the extracted image features and sequential text input[9].An LSTM layer, an embedding layer, and a fully connected output layer make up the architecture.

The process begins by feeding the image feature vector into a dense layer to match the embedding space of the captions. The LSTM's initial hidden state is this modified feature vector.

The LSTM predicts the subsequent word in the sequence after processing the previous word and the visual context at each time step[10]. The term with the highest probability is chosen after a softmax activation function has been used to create a probability distribution over the vocabulary. Until the final token, the procedure is repeated.("<end>") is produced after the created word is fed back into the LSTM. To improve learning, teacher forcing is used during training, where the ground-truth word is provided as input at each time step rather than the model's predicted word[13].During inference, captions are generated for new images using the trained CNN- LSTM architecture.

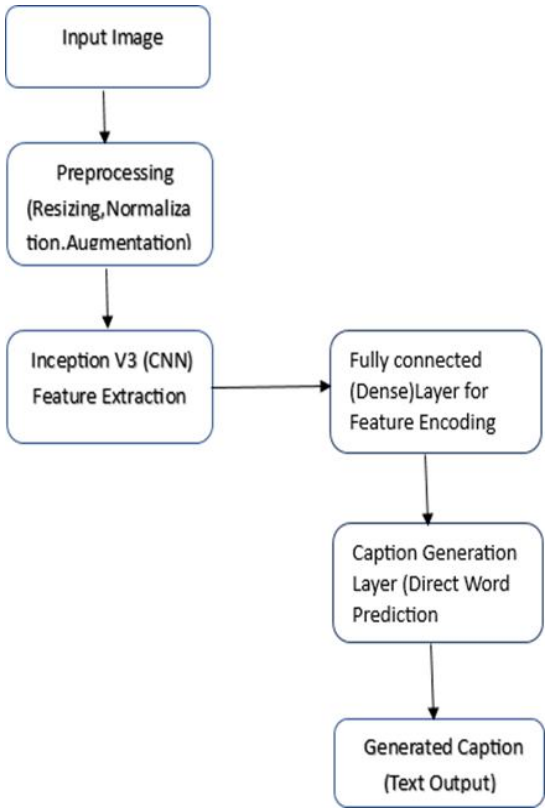


Fig 2. Block Diagram for Proposed Method

The image is first processed through InceptionV3 to extract feature vectors, which are then passed to the LSTM decoder along with the start token. The model iteratively predicts words, using previously generated words as input, until the end token is reached or a predefined maximum length is exceeded. To enhance the quality of generated captions, beam search decoding can be applied instead of a greedy approach. Unlike greedy decoding, which selects the most probable word at each step, beam search considers multiple candidate sequences simultaneously and selects the one with the highest overall probability, resulting in more coherent and contextually accurate captions.To evaluate the model's performance, several NLP-based metrics are used. The n-gram overlap between the reference and generated captions is measured by the BLEU Score (Bilingual Evaluation Understudy)[15].

To evaluate the quality of captions, METEOR (Metric for Evaluation of Translation with Explicit ORdering) takes into account word order, synonym matching, and stemming.. CIDEr (Consensus-based Image Description Evaluation) compares generated captions against multiple reference captions using term frequency- inverse document frequency (TF-IDF) weighting. ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) evaluates sequence-level recall and precision based on longest common subsequences[12].

These criteria drive future developments and help assess how effectively the algorithm produces captions that resemble those of a human. verall, the proposed approach effectively combines the strengths of CNN-based feature extraction and LSTM-based sequential modeling to create an automated image captioning system. By leveraging a pretrained InceptionV3 model and training on a large-scale dataset like MS-COCO[13], the system can generate meaningful and contextually relevant descriptions for diverse images. Further enhancements, such as attention mechanisms, transformer- based architectures, or reinforcement learning, can be explored to improve caption accuracy and fluency.

V. RESULTS

After defining and fitting the model. We trained our model for 50 epochs and it is 50epochs. It is observed that during the initial epochs of training the accuracyis very low and the captions generated are not much related to given test images. If we train the model for atleast 20 epochs then we have observed that the captions generated are some what related to the given test images. If the model is trained for 50 epochs we observe that the accuracy of the model increases and the captions generated are much related to the given teat images.

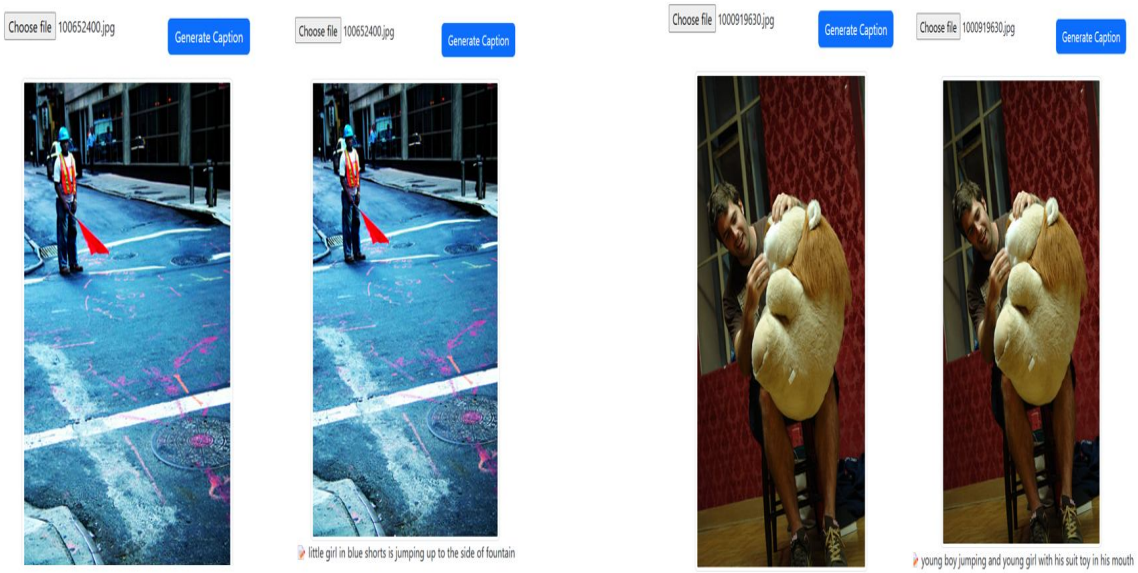


Fig 3. Existing System Outputs

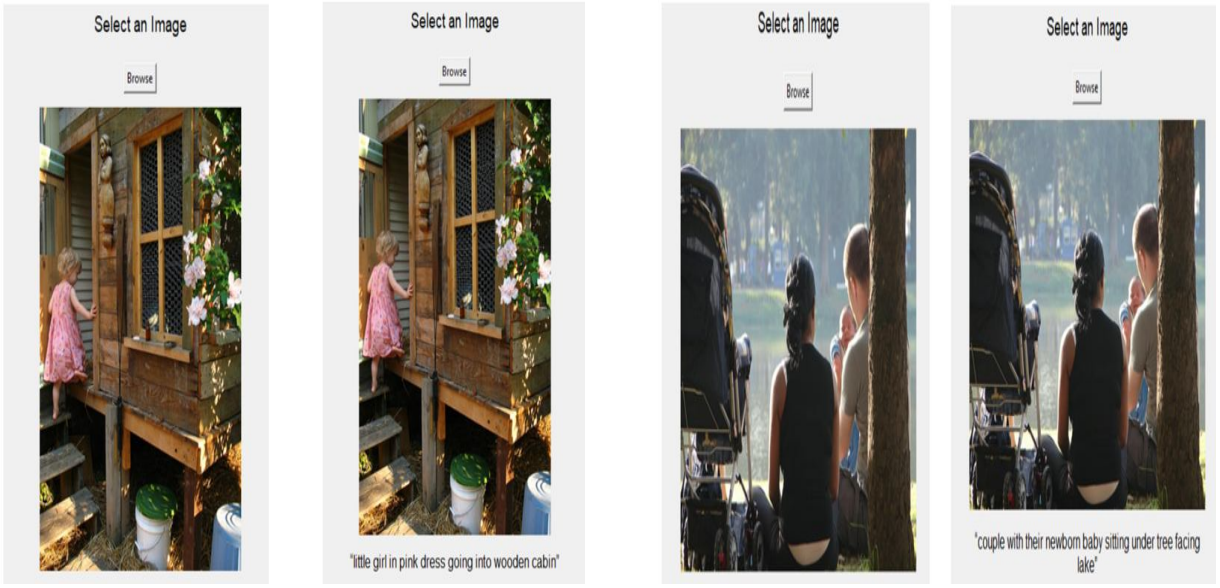


Fig 4. Proposed Method Outputs

VI. CONCLUSION

The automated image caption generator using deep learning effectively combines InceptionV3, a pretrained CNN, with an LSTM-based decoder to generate meaningful and contextually accurate captions. By leveraging the MS-COCO dataset, the system learns from diverse image-caption pairs, enabling it to describe new images with high relevance. Image processing techniques such as resizing, normalization, and feature extraction ensure that the model receives high-quality visual inputs, while beam search decoding enhances the fluency and coherence of generated captions.

Despite its success, the model faces challenges in describing complex scenes and unseen objects due to its reliance on training data. Future improvements, including attention mechanisms, transformer-based architectures, and domain-specific fine- tuning, could further enhance its accuracy and applicability. Overall, this project highlights the potential of deep learning- based image captioning in improving image accessibility, AI-driven content generation, and assistive technologies.

VII. FUTURE SCOPE

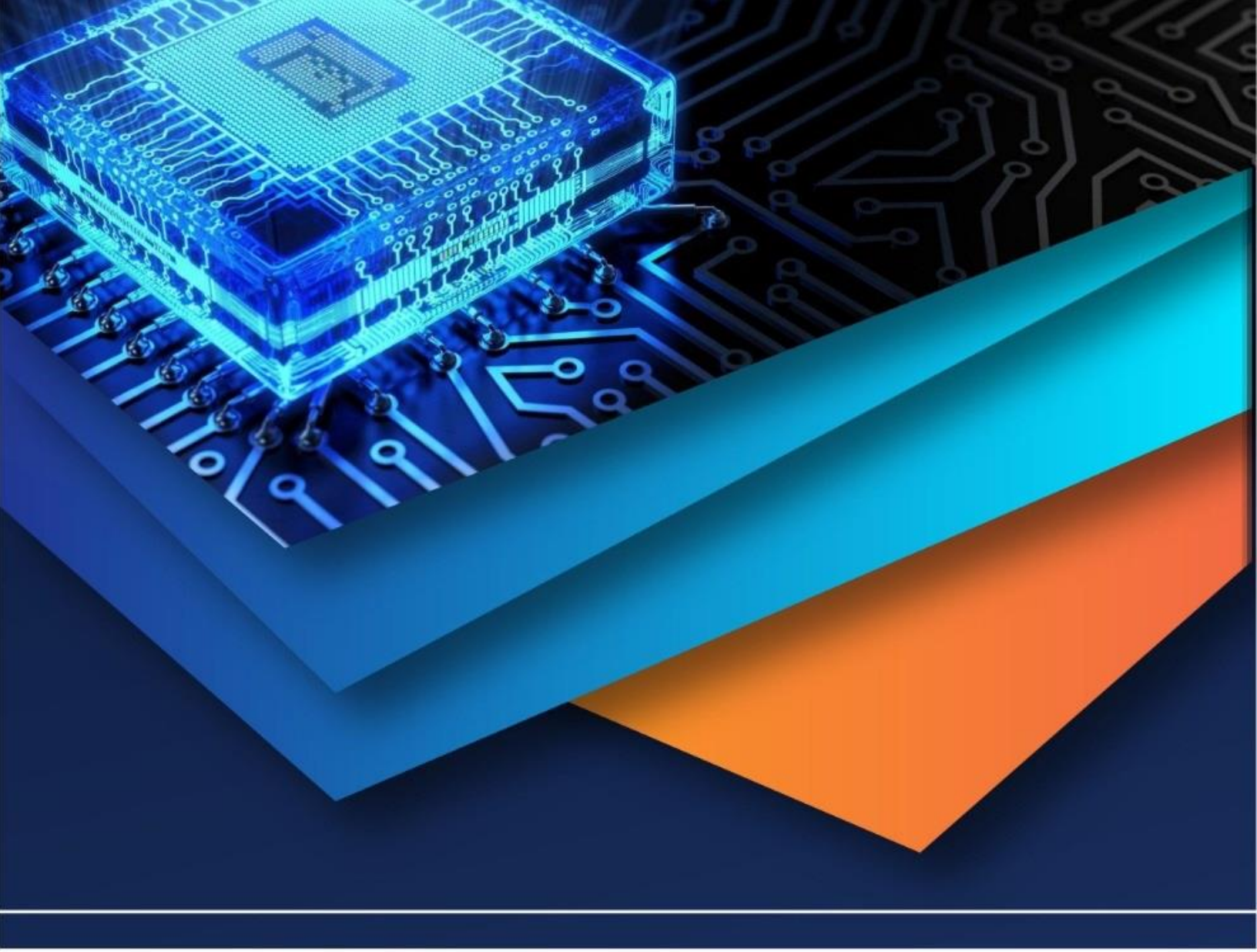
In the future, we plan to extend this model to create a complete image-to-speech conversion system, where the generated captions will be translated into speech, making the system more accessible, particularly for individuals with visual impairments. Aadditionally, as new architectures and techniques in deep learning continue to emerge, we aim to explore methods that can generate not only more accurate captions but also captions that exhibit a deeper understanding of context and object relationships within images.

VIII. ACKNOWLEDGMENT

At the outset, we thank God Almighty for making our endeavour a success. It is an immense pleasure to express our gratitude to our college management for kindly providing facilities in accomplishing our project work. We also express our gratitude to Dr. G. Srinivasa Rao, Principal, BWEC for providing us adequate facilities, ways and means by which we are able to complete this project work.We express our sincere gratitude to Mrs. B. Maha Lakshmi, M. Tech., (Ph.D.), HoD, Department of ECE, BWEC, for her support without which the successful completion of this project work would have not been possible. We express our immense pleasure and thanks to our guide Mrs. G. Krishna Veni, M. Tech., (Ph.D) Assistant Professor, Department of ECE, BWEC, for her constant co-operation and support.We also express heartfelt thanks to all our Classmates and Friends, for their constant co-operation and support throughout the project work.

REFERENCES

[1] Sun Chengjian, Songhao Zhu, Zhe Shi, "Image Annotation Via Deep Neural Network", Published in: 2015 14th IAPR International Conference on Machine Vision Applications (MVA), Pages 347- 350, DOI: 10.1109/MVA.2015.7153205.
[2] Venkatesh N. Murthy, Subhransu Maji,R.Manmatha, "Automatic Image Annotation using Deep Learning Representations", Published in: ICMR '15: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Pages 603-606, DOI: 10.1145/2671188.2749405
[3] Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator", Published in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Pages 3156-3164,DOI:
[4] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention", Published in: Proceedings of the 32nd International Conference on Machine Learning (ICML), Volume 37, Pages 2048-2057,2015.URL:https://proceedings.mlr.press/v37/xuc15.ht ml.
[5] Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation", Published in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Pages 311-318, 2002. DOI: 10.3115/1073083.1073135.
[6] Deep Learning Specialization, Offered by: deeplearning.ai, Available at: https://www.deeplearning.ai.
[7] TensorFlow: An end-to-end open-source machine learning platform, Available at: https://www.tensorflow.org.
[8] Gaurav, Pratistha Mathur, "Empirical Study of Image Captioning Models Using Various Deep Learning Encoders", Published in: Machine Learning and Computational Intelligence Techniques for Data Engineering, Lecture Notes in Electrical Engineering, Vol. 998, Pages 303-315, 2023
[9] Rashid Khan, Bingding Huang, Haseeb Hassan, Asim Zaman, Zhongfu Ye, "A Comparative Study of Pre-trained CNNs and GRU-Based Attention for Image Caption Generation", Published in: arXiv preprint arXiv:2310.07252, 2023.
[10] Aditya Bhattacharya, Eshwar Shamanna Girishekar, Padmakar Anil Deshpande, "Empirical Analysis of Image Caption Generation using Deep Learning", Published in: arXiv preprint arXiv:2105.09906,2021.
[11] Andrej Karpathy et al.(2017). Deep Visual-Semantic Alignments for Generating Image Descriptions". Proceedings - 4th International Conference on Computing, Communication Control and descriptions(CVPR), 1–4.
[12] Alec Radford et al. (2020)"Learning Transferable Visual Models From Natural Language Supervision" (CLIP): Advances, trends, applications, and datasets. The Visual Computer, 1–32.in axvix(2020)
[13] Aditya Ramesh et al. (2022) "Hierarchical Text-Conditional Image Generation with CLIP Latents",arXiv Neural Processing Letters, 50(1), 103–119. Georgios Barlas, Christos Veinidis, and AviArampatzis. What we see in a photograph:content selection for image captioning. TheVisual Computer, 37(6):1309–1326, 2021.
[14] Khaled Bayoudh, Raja Knani, Fay¸cal Ham-daoui, and Abdellatif Mtibaa. A surveyon deep multimodal learning for computervision: advances, trends, applications, anddatasets. The Visual Computer, pages 1–32,2021.
[15] Rajarshi Biswas, Michael Barz, and DanielSonntag. Towards explanatory interac-tive image captioning using top-downand bottom-up features, beam search andre-ranking. KI-K¨unstliche Intelligenz,34(4):571–584, 2020.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY