IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Image Caption Generation

Prof. Kriti Sachdeva[1], Mrinal Pandit[2], Nikhil Patil[3], Abhishek More [4] and Ronit Thakur[5]

Abstract: The Image Caption Generator's job is to give subtitles for the supplied photos. A natural language is created by extracting and transforming the semantic content of the image. Capturing entails a difficult procedure that combines picture processing and computer vision. People, animals, and things must all be recognized by the system, and relationships must be made between them.

This paper aims to detect, recognize, and generate worthwhile captions for a given image using deep learning. A Regional Object Detector (RODe) is used for the detection, recognition, and generation of captions**.** The practice of creating textual descriptions of an image by utilizing computer vision and natural language processing methods is known as image captioning. For the newer models to perform better, we used deep learning methods for this purpose. But these models can't tell you which things in a picture are more essential than others, or they can't tell you why certain phrases were used in the captions.

Index Terms **- Deep learning, feature extraction, thresholding, image segment, image captioning.**

1. INTRODUCTION

Significant progress has been made in the field of picture understanding and interpretation in recent years as a result of the combination of computer vision and natural language processing. Creating picture caption generation models is an intriguing use of this multidisciplinary technique. To provide descriptive and contextually appropriate captions for photos, this paper explores and implements a complex architecture.

The term "image caption generations" describes how captions or descriptions for images change over time. The accuracy, descriptiveness, and accessibility of picture captions have increased with the development of technology and artificial intelligence. This development helps people who are blind or visually impaired, optimizes content for search engines, and enhances user interfaces on a variety of digital platforms.

2. MOTIVATION

The inspiration driving this examination lies in the developing requirement for machines to speak with people more instinctively and regularly. The generation of image captions can be used for a variety of things, like making content easier to index and search and making it more accessible to people with visual impairments. In addition, it contributes to the burgeoning field of multimodal AI, in which machines can comprehend and describe the world in a manner that is more human-like thanks to the convergence of visual and textual information:

1) Enhancing Accessibility: Contributing to the development of inclusive technology is one of the main drives. Image caption generation can create a more inclusive digital environment by offering written explanations of visual material, which can increase accessibility for those with visual impairments.

2) Content Indexing and Retrieval: Nowadays, with so many digital archives, effective content indexing and retrieval is critical. A sophisticated picture captioning system can help with better visual data organization and retrieval, which can lead to enhanced search capabilities and user experiences.

3) Human-Machine Communication: The desire to develop AI systems that can converse with people in a way that is more in line with human thought processes is what drives the effort. By bridging the gap between

natural language and the visual world, picture caption generation allows machines to explain what they understand from images in a way that is understandable and accessible to people.

4) Applications in Diverse Domains: This research is motivated by the adaptability of image captioning applications in different sectors. The possible uses are numerous and range from helping medical professionals analyze diagnostic images to adding descriptive visual content to educational materials.

3. REVIEW OF LITERATURE

"**Explainable Image Caption Generator Using Attention and BayesianInference" Author: Seung-Ho Han and Ho-Jin Choi** The act of creating textual descriptions for a picture is known as image captioning, and it involves computer vision and natural language processing algorithms. To boost performance, recent models have applied deep learning approaches to this task. Nevertheless, these models are unable to identify objects in a picture that are more significant than others or provide an explanation for the word choices used during the caption generation process[1]

"Domain-SpecificImageCaptionGeneratorwithSemanticOntology" Author: Seung-Ho Han and Ho-JinChoi The act of creating textual descriptions for a picture is known as image captioning, and it involves computer vision and natural language processing algorithms. To boost performance, recent models have applied deep learning approaches to this task. Unfortunately, because previous methods employ publicly available datasets like MSCOCO, which covers broad photos, these models are unable to fully utilize information provided in a given image, such as object and attribute, or to produce a caption appropriate to a certain domain.[2]

"Image Captioning with Generative Adversarial Network" Author: Soheyla Amirian, Khaled Rasheed, Thiab R. Taha, Hamid R.Arabnia There is significant overlap in the approaches used by automatic picture annotation, automatic image tagging, and image linguistic indexing functions. In this work, all variations of these functions are referred to by the general term "image captioning." The act of automatically producing captioning—that is, creating phrases that explain the contents of a picture—is known as image captioning [3]

"Explainable AI(XAI) approach to image captioning" Author: Seung-Ho-Han, Min-Su Kwan, Ho-Jin Chai The method for picture captioning described in this article is called eXplainable AI (XAI). Deep learning methods have been applied heavily to this issue recently, with comparatively strong results. However, because of deep learning's "black-box" model, current methods can't offer hints as to why particular words have been chosen when creating captions for particular photos, which occasionally results in the creation of ridiculous captions. This article suggests an explainable picture captioning model as a solution to this issue. It creates a visual connection between a word (or phrase) in the created sentence and the specific area of an item (or concept) in the given image. Two datasets, MSCOCO and Flickr30K, were used to evaluate the model. The quantitative and qualitative findings show the effectiveness of the proposed model.[7]

4. Purpose:

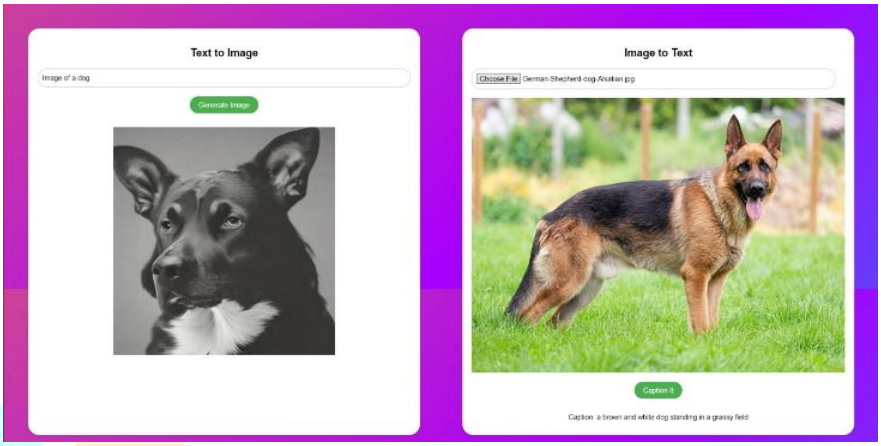
To make visual material more accessible to a larger audience, including people with visual impairments, image caption generation aims to produce meaningful and helpful explanations for pictures. The goals of this technology are to help organize material, optimize search engine results, and improve user experiences. It facilitates improved human-computer interaction and helps with activities like content retrieval and recommendation by creating captions, which also helps with content indexing and helps computers grasp the context of pictures

5. RESULTS

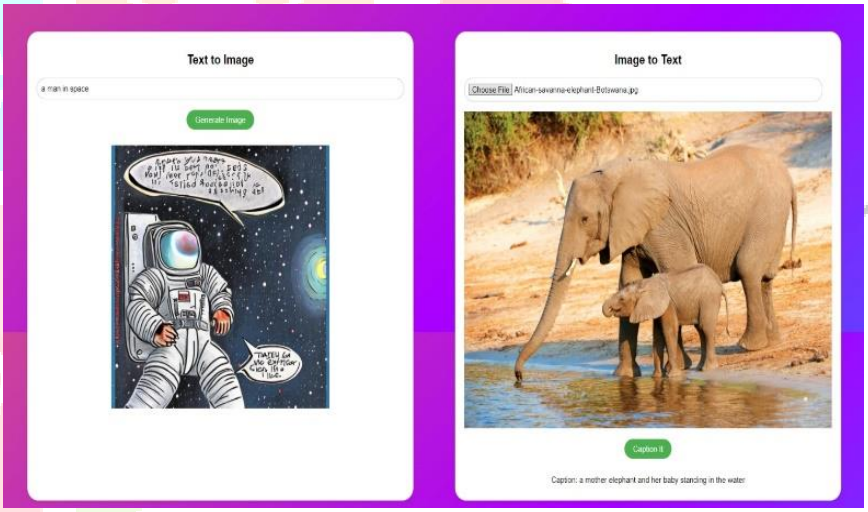
Working Models

1) Animal Category:

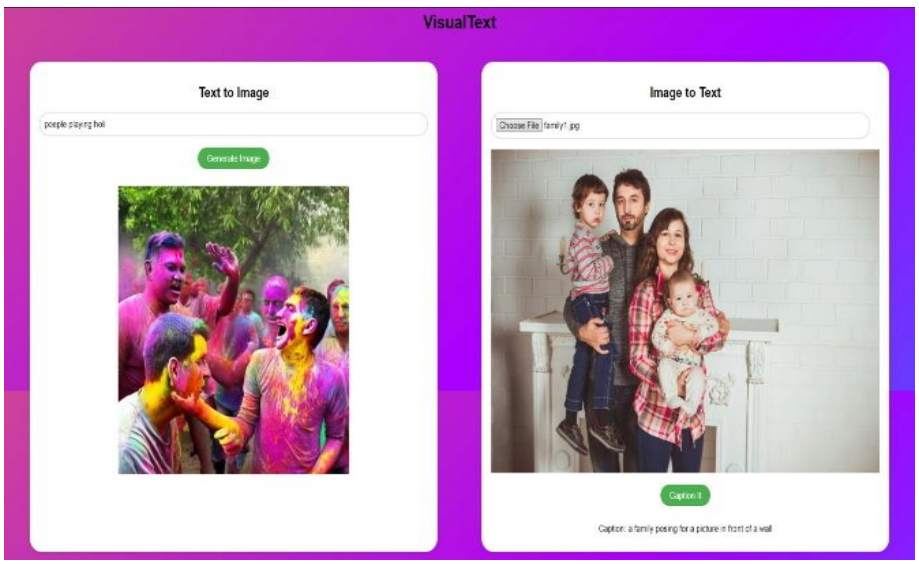
- **Dog**



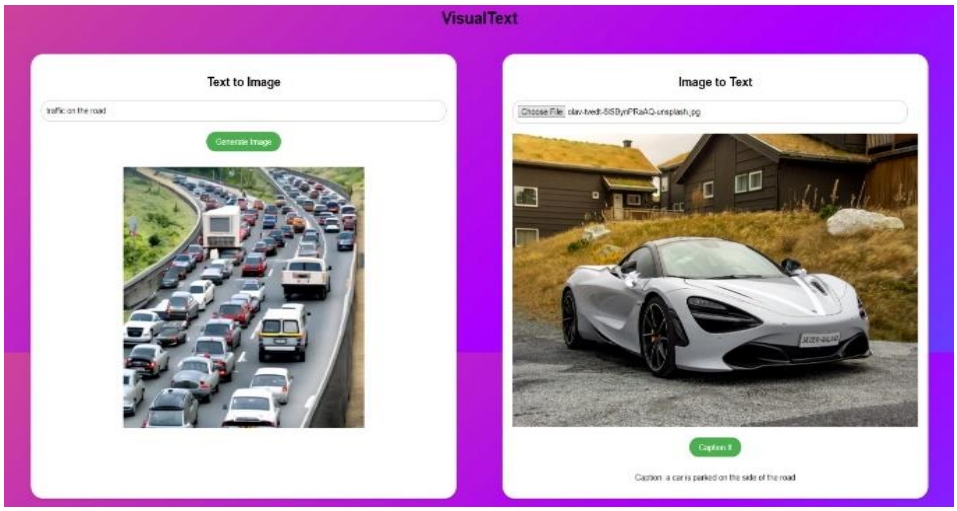
- Elephant



2) Human



3) Car



6. FUTURE SCOPE

The potential for picture caption generation is enormous given the continuous progress being made in computer vision and artificial
intelligence. We can expect more advancements and applications in the following areas:

1) **Improved Model Architectures:** More advanced model architectures that more accurately represent the subtleties of both visual and textual data are probably going to be produced by future research. This might entail the inclusion of transformer designs, more sophisticated attention processes, or creative arrangements of neural network elements.

2) **Multimodal Models with Audio and Video:** In the future, picture captioning systems might accommodate dynamic content like videos in addition to static photographs. Combining audio data with textual and visual modalities may result in a more thorough and sophisticated understanding, making it possible for AI systems to produce subtitles for a wider variety of multimedia content.

3) **Fine-Grained Understanding:** A shift towards fine-grained comprehension, in which models can explain not just the items in a picture but also their relationships, the emotions they express, or minute contextual data, could be a step forward in image captioning. This could result in descriptions that are more complex and human.

4) **Domain-Specific Applications:** One possible approach is to customize image captioning models for particular domains, including industrial applications, satellite photography, or medical imaging. Better and more specialized descriptions may be offered by customized models that have been trained on datasets unique to a given topic.

5) **Robustness and Ethical Considerations:** Subsequent investigations will probably concentrate on enhancing the resilience of picture captioning models, guaranteeing their efficacy in a variety of cultural contexts and demographic groups. To guarantee impartial and equitable outcomes, ethical issues like prejudice in caption creation must also be addressed.

6) **Real-Time Applications:** Exciting opportunities arise when picture captioning is incorporated into real-time applications like augmented reality (AR) or live video streaming. In dynamic situations, artificial intelligence (AI) systems have the potential to produce captions instantly and pertinently.

7) **Cross-Modal Learning:** Developments in cross-modal learning, in which models are trained to comprehend and produce content in many modalities (e.g., text and images), may result in more adaptable and comprehensive AI systems that can manage a variety of data kinds.

8) **Human-AI Collaboration**: In the future, there might be more cooperation between AI and humans when it comes to captioning. Interactive tools that let users improve or direct the captioning process could improve the originality and personalization of the generated captions.

7. CONCLUSION

In conclusion, this research presents a deep-learning approach employing neural networks to generate image captions. The suggested method was tested on a dataset. Compared to other image caption generators now in use, the suggested deep learning technology generated captions with greater descriptive meaning. For more precise captions, a hybrid model of an image caption generator might be created in the future. a model that analyzes items and provides an explanation for the word choices it makes when creating a caption for a given image. To do that, we developed an explanation module that produces an image–sentence

ACKNOWLEDGMENT

We would like to extend our heartfelt gratitude to our project guide Prof. Kriti Sachdeva, whose guidance and technical expertise have been instrumental in the successful completion ofthis research. Her profound knowledge and unwavering support have been a constant source of inspiration for us. We are also immensely grateful to the Head of the Department Dr. Prof. Mrs.Sarita Deshpande, whose leadership and vision have significantly contributed to our work. Her continuous encouragement and faith in our abilities have been pivotal in shaping this research. Their collective wisdom and support havebeen invaluable to us, and we thank them for their trust and guidance.

REFERENCES

1. K. He, G. Gkioxari, P. Dolla´r, and R. Girshick, “Mask r-cnn,” in Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
2. J. Redmon, S. Divvala, Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection”, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.
3. D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprintarXiv:1409.0473, 2014.
4. R. Kiros, R. Salakhutdinov, and R. S. Zemel, “Unifying visual-semantic embeddings with multimodal neural language models,” arXiv preprint arXiv:1411.2539, 2014.
5. “Explainable Image Caption Generator Using Attention and Bayesian Inference”, Institute of Electrical and Electronics Engineers (IEEE), 19245459,12-14-December,2018https://ieeexplore.ieee.org/document/8947893
6. “Image Captioning with Generative Adversarial Network”, Institute of Electrical and Electronics Engineers (IEEE), 19535763, 05-07 December 2019, https://ieeexplore.ieee.org/document/9071372
7. “Domain Specific Image Caption Generator with Semantic Ontology”, Institute of Electrical and Electronics Engineers (IEEE), 19534889, 19-22-February 2020, https://ieeexplore.ieee.org/document/9070680
8. S. ALBAWI and T. A. MOHAMMED, "Understanding of a Convolutional Neural Network," in ICET, Antalya, 2017
9. S. Hochreiter, "LONG SHORT-TERM MEMORY," NeuralComputation, December 1997

10. O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "A NeuralImage Caption Generator," CVPR 2015 OpenAccessRepository, vol. Xiv, 17 November 2014.

11. D. S. Whitehead, L. Huang, H., and S.-F. Chang, "Entity-aware Image CaptionGeneration,"inEmpiricalMethodsin Natural Language Processing, Brussels, 2018.

12. C. Elamri and T. Planque, "Automated Neural Image Caption Generator for Visually Impaired People, "California, 2016.

13. G. Ding, M. Chen, S. Zhao, H. Chen, J. Han and Q. Liu, "Neural Image Caption Generation with WeightedTraining and Reference," Cognitive Computation, 08 August 2018.

14. J. Chen, W. Dong, and M. Li, "Image Caption Generator Based On Deep Neural Networks," March 2018.

15. S. Bai and S. An,"A Survey on Automatic ImageCaption Generation," Neurocomputing, 13 April 2018.

16. J. Hessel, N. Savva and M. J. Wilber, "Image Representations and New Domains in Neural Image Captioning," ACL Anthology, vol. Proceedings of the Fourth Workshop on Vision and Language, p. 29–39, 18 September 2015.