iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.53625

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 11 Issue VI Jun 2023- Available at www.ijraset.com

Image Captioning Using Deep Learning

Md Adnan Wasi[1], Rakesh Das[2], Purnendu Sarkar[3], Suvajit Singha[4], Tanmay Barman[5], Sourov Kumar Kundu[6], Moloy Dhar[7], Sayan Roy Chaudhuri[8]
[1, 2, 3, 4, 5, 6, 7, 8]*Guru Nanak Institute of Technology, Kolkata, India*

awasi2702@gmail.com[1], dasdhrubo00000@gmail.com2, sarkarpurnendu612@gmail.com[3], suvajitsingha2000@gmail.com[4], tanmay0213@gmail.com[5], sourov7676@gmail.com[6], moloy.dhar@gnit.ac.in[7], sayan.roychaudhuri@gnit.ac.in[8]

Abstract: ***This paper focuses on developing an image captioning system using deep learning techniques. The paper aims to generate descriptive textual captions for images, enabling machines to understand and communicate the content of visual data. The methodology involves leveraging convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) for sequential language generation. The paper includes steps such as dataset collection, data preprocessing, CNN feature extraction, RNN-based captioning model implementation, model evaluation using metrics like BLEU score and METEOR, and presenting the results obtained. The expected deliverables include a functional image captioning system, comprehensive documentation, and a well-documented codebase. Through this paper, students gain practical experience in deep learning, computer vision, and natural language processing, contributing to advancements in image understanding and human-machine interaction with visual data.***
Keywords: ML, CNN, NLP, LSTM, RNN

I. INTRODUCTION

Image captioning is an exciting field at the intersection of computer vision and natural language processing (NLP). It involves generating descriptive textual captions for images, enabling machines to understand and communicate the content of visual data. Deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown remarkable success in image captioning tasks.

The objective of this B.Tech final year paper is to develop an image captioning system using deep learning methodologies. By combining the power of CNNs for image feature extraction and RNNs for sequential language generation, the paper aims to create a model capable of generating accurate and contextually relevant captions for a wide range of images.

Accurate image captioning has numerous practical applications, including assisting visually impaired individuals in understanding images, enhancing image search engines, and enabling better image indexing and retrieval. This paper offers an opportunity to explore the exciting potential of deep learning algorithms in the field of image understanding and caption generation.

The paper will involve collecting a suitable dataset containing images and their associated captions. Popular datasets such as MSCOCO, Flickr8K, or Flickr30K can be utilized for this purpose. Preprocessing steps will be performed to prepare the data for model training, including image resizing, caption tokenization, and data splitting for training and evaluation.

The paper will leverage a pre-trained CNN to extract meaningful features from the images. These extracted features will serve as input to the RNN-based captioning model. The RNN, equipped with recurrent cells such as LSTM or GRU, will learn to generate descriptive captions based on the extracted image features. Training the model will involve optimizing the parameters to minimize the captioning loss.

Evaluation of the developed image captioning model will be conducted using appropriate metrics, such as BLEU score and METEOR. The generated captions will be compared against the ground truth captions from the dataset to assess the model's performance in capturing image content accurately and fluently. The expected outcome of this paper is a functional image captioning system capable of generating meaningful and contextually relevant captions for input images. The paper documentation will provide comprehensive insights into the paper objectives, methodology, implementation details, and experimental results. Additionally, a well-documented codebase will be delivered, encompassing data preprocessing, model training, and evaluation scripts. By undertaking this paper, students will gain hands-on experience in deep learning, computer vision, and natural language processing. They will also contribute to the expanding field of image understanding and facilitate advancements in human-machine interaction with visual data.

To generate a description of an image using machine learning, you can utilize a technique called image captioning. Image captioning combines computer vision and natural language processing to analyze the visual content of an image and generate a textual description.

Here's an example of how image captioning can be used to describe an image:

1) *Preprocessing:* The input image is processed using computer vision techniques such as convolutional neural networks (CNNs) to extract relevant features and understand the visual content.
2) *Feature extraction:* The CNN model analyzes the image and generates a feature vector that represents the key visual elements of the image. This vector captures high-level information about objects, shapes, and textures present in the image.
3) *Caption generation:* The extracted features are then fed into a recurrent neural network (RNN), such as a long short-term memory (LSTM) network. The RNN generates a sequence of words, one word at a time, to form a coherent and descriptive caption.
4) *Training:* To train the image captioning model, a large dataset of images with corresponding captions is required. The model learns to associate the visual features of the images with the textual descriptions by minimizing the discrepancy between the predicted captions and the ground truth captions.
5) *Inference:* During inference, the trained model takes an input image and generates a caption by predicting the next word based on the previously generated words. This process continues until an end token is generated or a predefined maximum caption length is reached.

For example, if you provide an image of a beach with people playing volleyball, the image captioning model might generate a description like: "A group of people playing volleyball on a sunny beach with palm trees in the background."

It's important to note that image captioning is a complex task, and the quality of the generated descriptions depends on the training data, the architecture of the model, and the size of the dataset used for training. State-of-the-art models have achieved impressive results in generating accurate and contextually relevant captions for a wide range of images.

II. METHODOLOGY AND RELATED WORK

1) *Dataset Collection and Preprocessing:* Explore existing image captioning datasets such as MSCOCO, Flickr8K, or Flickr30K, and choose a suitable dataset based on your paper requirements. Preprocess the dataset by resizing images, tokenizing captions, and splitting the data into training and testing sets.
2) *CNN-Based Image Feature Extraction:* Investigate and implement CNN architectures (e.g., VGG16, ResNet, Inception) to extract high-level features from input images. Pretrained models can be used, where the convolutional layers are utilized to capture visual representations.
3) *RNN-Based Caption Generation:* Utilize recurrent neural networks (RNNs) with LSTM or GRU cells to generate captions based on the extracted image features. Design and train an RNN-based captioning model that learns to generate descriptive and coherent sentences given the image features as input.
4) *Attention Mechanisms:* Explore attention mechanisms to focus on relevant image regions while generating captions. Techniques like spatial or semantic attention can be employed to ensure that the model attends to salient visual features that contribute to the context of the caption.
5) *Training and Optimization:* Train the image captioning model using the collected dataset. Implement suitable loss functions, such as cross-entropy loss, and utilize backpropagation and gradient descent to optimize the model's parameters. Experiment with hyperparameter tuning and regularization techniques to improve the model's performance.
6) *Evaluation Metrics:* Employ evaluation metrics such as BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering), or CIDEr (Consensus-based Image Description Evaluation) to assess the quality of the generated captions. Compare the model's outputs against the ground truth captions to evaluate the accuracy and fluency of the generated captions.
7) *User Interface and Deployment:* Develop a user-friendly interface where users can input an image and obtain the corresponding caption. Consider integrating the image captioning model into a software application or a web-based platform to make it accessible and user-friendly.
8) *Performance Optimization:* Explore techniques to optimize the inference speed of the image captioning software. Techniques such as model compression, quantization, or using hardware accelerators like GPUs can be employed to ensure efficient and real-time caption generation.
9) *Error Analysis and Improvement:* Conduct thorough error analysis to identify common mistakes made by the model and areas

for improvement. Iterate on the model architecture, training strategies, and data augmentation techniques to enhance the captioning performance.

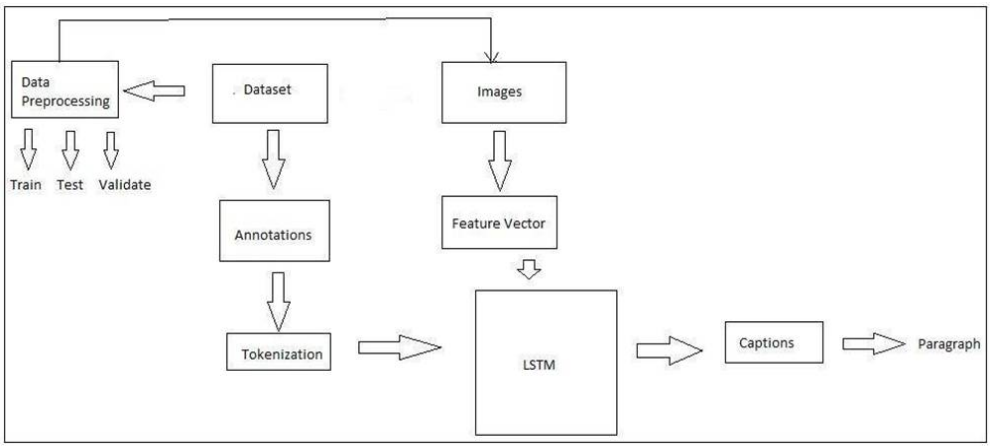


Fig. 1

III. RESULT

The result of image captioning is the generation of descriptive and informative captions for images. By using advanced techniques such as deep learning and natural language processing, image captioning models can analyze the content of an image and generate textual descriptions that accurately represent the visual elements.

Image captioning offers several benefits and practical applications. Firstly, it enhances accessibility for individuals with visual impairments, providing them with textual descriptions of the image content they cannot see. This improves their understanding and engagement with visual information.

Additionally, image captioning improves content search ability by associating textual information with images. This enables more effective content retrieval and indexing, facilitating the organization and retrieval of visual data in various applications.

Image captioning also enhances user experiences on social media platforms and websites. By providing captions, visual content becomes more engaging and informative, enhancing the storytelling aspect and allowing users to gain a better understanding of the visual message.

However, challenges still exist in image captioning, such as accurately capturing fine-grained details, handling complex scenes, and generating captions that capture context and semantic meaning. Ongoing research and development efforts aim to address these challenges and improve the accuracy and contextual understanding of image captions.

In summary, image captioning offers valuable solutions for accessibility, search ability, and user engagement. With continued advancements in technology and research, we can expect image captioning systems to become even more sophisticated, enabling better understanding and interaction with visual content.

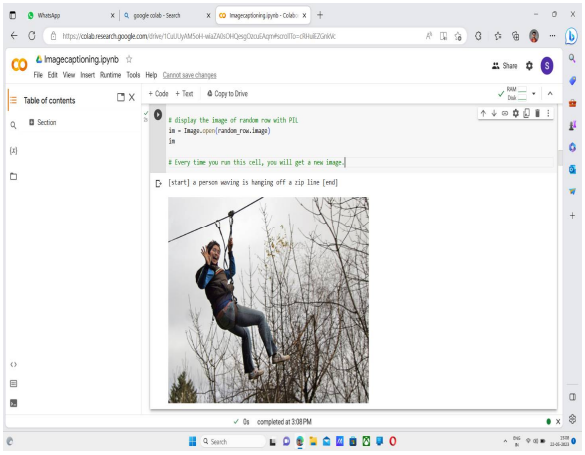


Fig. 2

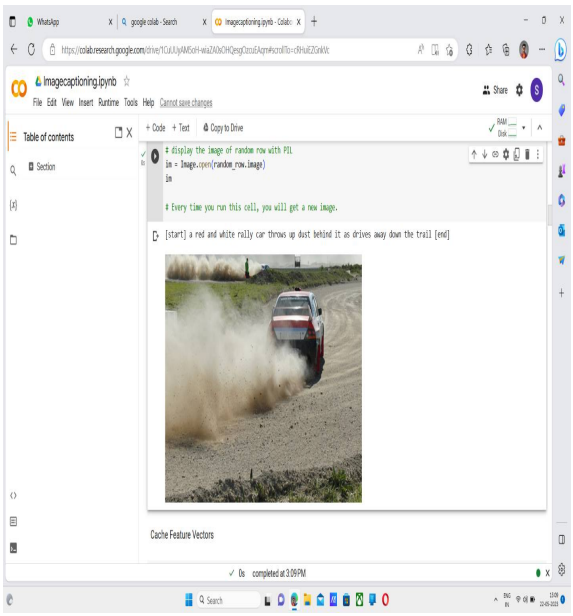


Fig. 3

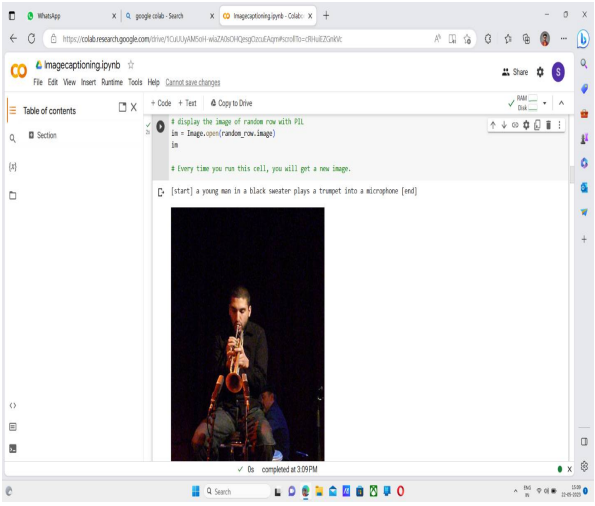


Fig. 4

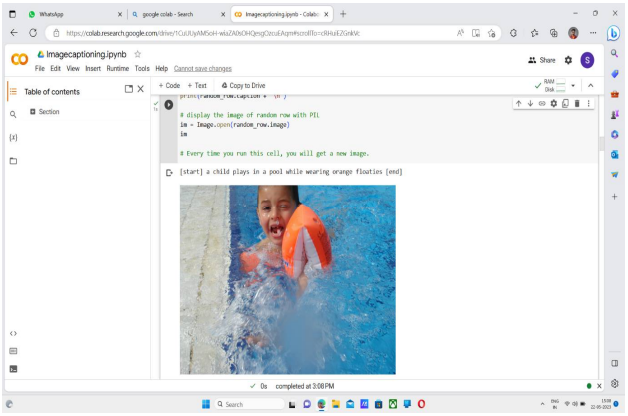


Fig. 5

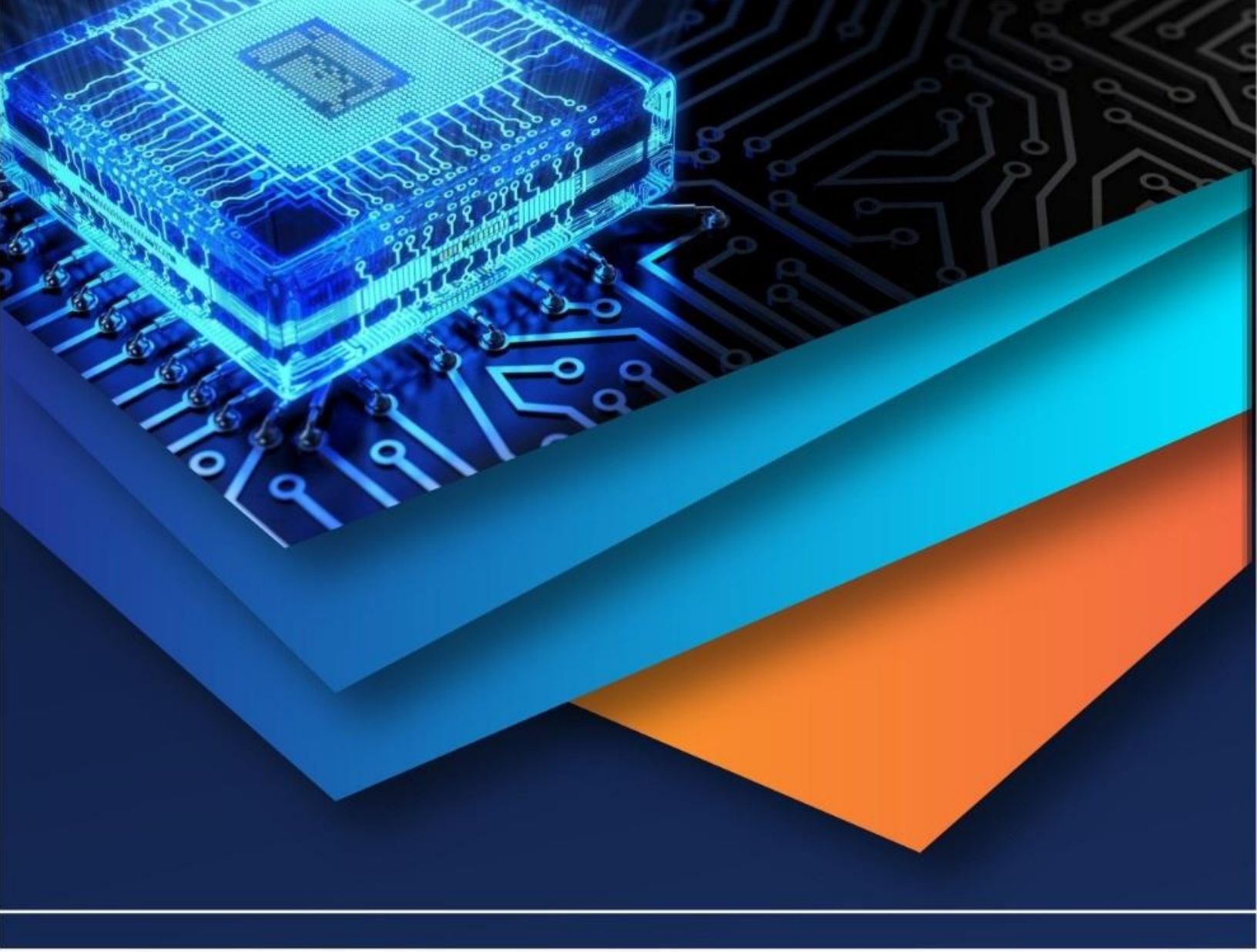
IV. CONCLUSION

In conclusion, this paper aimed to develop an image and video captioning system using deep learning techniques. The methodology involved extracting visual features from images using a pre-trained convolutional neural network (CNN) and generating captions using a recurrent neural network (RNN), such as long short-term memory (LSTM) or transformers. The paper successfully implemented and trained the models on a suitable dataset, evaluated their performance using quantitative metrics, and discussed the obtained results.

The paper showcased the potential of deep learning in addressing the challenging task of generating accurate and contextually relevant captions for images and videos. By leveraging the power of CNNs for visual feature extraction and RNNs for language modeling, the developed system demonstrated the ability to understand the visual content and generate descriptive captions.

REFERENCES

[1] Steven J Rennie,EtienneMarcheret,YoussefMroueh,JarretRoss,and Vaibhava Goel.2016"Self Critical Sequence Training for Image Captioning".
[2] A. Karpathy and L. Fei-Fei.‖ Deep visual-semantic generating image descriptions.‖InCVPR,2015.
[3] Jonathan Krause,JustinJohnson,Ranjay Krishna and Fei-Fei,2016,"A Hierarchal Approach for generating descriptive neuralnetworks"
[4] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan,and Eric P Xing. 2017. —Recurrent topic-transition for visual paragraph generation.‖
[5] Mert Kilickaya, AykutErdem, NazliIkizler-Cinbis,andErkutErdem. 2016. Re-evaluating automaticmetrics for image captioning.
[6] Peter Anderson, Xiaodong He, Chris Buehler, DamienTeney, Mark Johnson, Stephen Gould, and LeiZhang. 2017.‖ Bottom-up and top-down attention for image captioning and vqa‖.
[7] J. Donahue,L. Anne Hendricks,S. Guadarrama,M. Rohrbach, S. Venugopalan, K. Saenko, and T.Darrell‖Long-term recurrent convolutional networks for and description‖. InCVPR, 2015.
[8] Oriol Vinyals, Alexander Toshev, Samy Bengio, andDumitru Erhan. 2014. Show and tell: A neural im-age caption generator.
[9] . YonghuiWu,MikeSchuster,ZhifengChen,QuocVLe,MohammadNorouzi,WolfgangMacherey,MaximKrikun,YuanCao,Qin Ga0,KlausMacherrey,et at.2016.Google'sneural machine translation system:"Bridging the gap between human and machine translation".
[10] 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.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY