iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** V **Month of publication:** May 2024

DOI: https://doi.org/10.22214/ijraset.2024.62269

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 12 Issue V May 2024- Available at www.ijraset.com

Image Caption Prediction Using Deep Learning

M.LalithaKeerthana[1], N.Leela Vallabha[2], P.Likitha[3], T.Lohith[4], Prof C.M.Preethi[5]
School Of Engineering B. Tech Computer Science- AIML Malla Reddy University India

Abstract: ***In the era of rapidly growing digital content, the automatic generation of image captions has gain significant attention in the field of deep learning and computer vision. Our proposed model leverages deep learning techniques, including convolutional neural networks (CNNs) for image feature extraction and recurrent neural networks (RNNs) with attention mechanisms for caption generation. To evaluate the performance of our model, we employ a comprehensive dataset containing a diverse range of images and corresponding human-generated captions. In conclusion, our deep learning-based image captioning project offers a promising solution to the challenge of automatically generating meaningful and contextually accurate descriptions for images. The potential applications of the technology are vast, and we anticipate that it will continue to various fields, including artificial intelligence, accessibility, content creation, content management and publishing across various social media platform.***

I. INTRODUCTION

With reference to the significant advancements in the field of technology, there has been a lot of digitalization and day by day, the revolution of the digital era has been impacting in an immense way. Data has become a crucial part in the field of technology. The data can be any text, images, audio, video, statistics from any institution or organization. As years pass on, a lot of emerging technologies have driven into the market and creating a positive impact. One such finest and top most emerging technology is AI. Artificial Intelligence is now at the heart of innovation economy and thus the base for this project is also the same. In the recent past a field of AI namely Deep Learning has turned a lot of heads due to its impressive results in terms of accuracy when compared to the already existing Machine learning algorithms. The task of being able to generate a meaningful sentence from an image is a difficult task but can have great impact, for instance helping the visually impaired to have a better understanding of images. With the great advancement in computing power and with the availability of huge datasets, building models that can generate captions for an image has become possible.

II. LITERATURE SURVEY

A written description must be provided for a given image as part of the difficult artificial intelligence challenge known as caption creation. It takes both approaches from computer vision to understand the content of the image and a language model from the field of natural language processing to transfer the comprehension of the image into words in the appropriate order. On applications of this problem, deep learning techniques recently produced state-of-the-art results. Deep learning techniques have delivered cutting-edge outcomes for caption generating issues. The most amazing aspect of these methods is that, rather than requiring complex data preparation or a pipeline of specially created models, a single end-to-end model can be developed to pAredict a caption given a photo. RNNs are now quite potent especially for modelling sequential data.

III. PROBLEM DEFINATION

Most of the people spend about hours deciding about what to write as a caption for an image. A picture is incomplete without a good caption to go with it. The problem introduces a captioning task, which requires computer vision system to both localize and describe salient regions in images in natural language. The image captioning task generalizes object detection when the descriptions consist of a single word. Given a set of images and prior knowledge about the content find the correct semantic label for the entire image(s). On the other hand, humans are able to easily describe the environments they are in. Given a picture, it's natural for a person to explain an immense amount of details about this image with a fast glance. Although great development has been made in computer vision, tasks such as recognizing an object, action classification, image classification, attribute classification and scene recognition are possible but it is a relatively new task to let a computer describe an image that is forwarded to it in the form of a human-like sentence.

©IJRASET: All Rights are Reserved

IV. OBJECTIVE OF PROJECT

The objective for image caption prediction is to generate a natural language description of an image. This is typically achieved by using computer vision techniques to extract features from the image, and natural language processing techniques to generate the description.

The scope of image caption prediction includes a wide range of applications in fields such as computer vision, natural language processing, and human-computer interaction. Some examples of the potential applications of image caption prediction include:

1) Image search and retrieval: Image caption prediction can be used to generate descriptions of images, which can then be used to search for and retrieve relevant images based on the content of the description. Assistive technology: Image caption prediction can be used to provide descriptions of images for people with visual impairments, allowing them to better understand and interact with visual content.
2) Social media: Image caption prediction can be used to automatically generate captions for images shared on social media platforms, making it easier for users to quickly and easily share and understand visual content. E-commerce: Image caption prediction can be used to generate descriptions of products for online retailers, making it easier for customers to search for and find the products they are looking for. **Surveillance and security:** Image caption prediction can be used to automatically generate descriptions of surveillance footage, making it easier for security personnel to quickly and accurately identify and respond to potential security threats. Overall, the scope of image caption prediction is quite broad and has the potential to be applied in many different domains

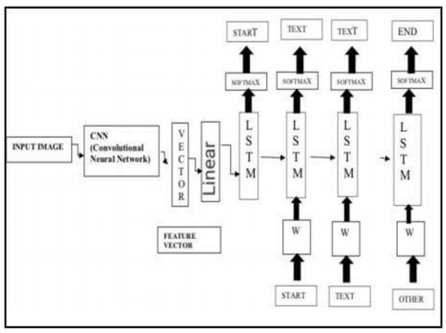
V. METHODOLOGY

Most of the people spend about hours deciding about what to write as a caption for an image. A picture is incomplete without a good caption to go with it. The problem introduces a captioning task, which requires a computer vision system to both localize and describe salient regions in images in natural language. The image captioning task generalizes object 1. detection when the descriptions consist of a single word. Given a set of images and prior knowledge about the content find the correct semantic label for the entire image(s). On the other hand, humans are able to easily describe the environments they are in Given a picture, it's natural for a person to explain an immense amount of details about this image with a fast glance. Although great development has been made in computer vision, tasks such as recognizing an object, action classification, image classification, attribute classification and scene recognition are possible but it is a relatively new task to let a computer describe an image that is forwarded to it in the form of a human-like sentence

VI. MODULES

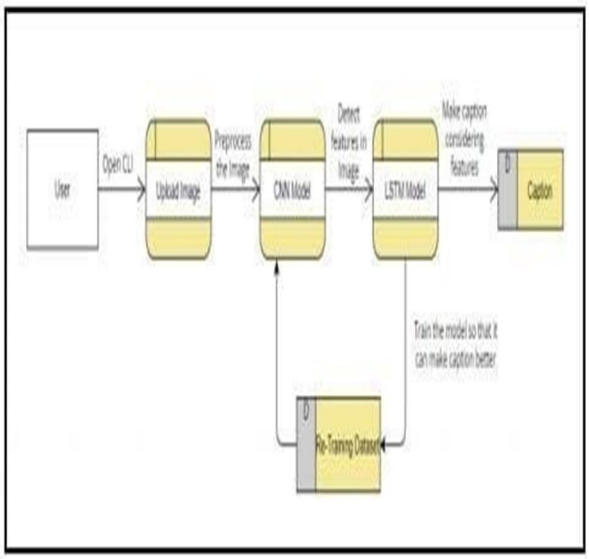
1) *Data Preprocessing:* The dataset is loaded and checked for duplicates. The 'Time' and 'Amount' columns are standardized using StandardScaler. Non-scaling 2. columns ('Time' and 'Amount') are dropped from the DataFrame Class imbalance is addressed using RandomUnderSampler from imblearn library. The dataset is split into training and testing sets.
2) *Model Development Module:* Logistic Regression model is initialized. The model is trained on the training data.
3) *Model Evaluation:* Predictions are made on the test set. Evaluation metrics such as accuracy, F1 score, precision, recall, and classification report are calculated. A confusion matrix is generated to assess model performance.
4) *Model Saving:* Evaluates the performance of the trained model on separate Code.

A. Architecture



ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

B. Flowchart



VII. METHODS AND ALGORITHMS

A. Convolutional Neural Networks (CNNs)

Convolutional Neural Network (CNN) is a Deep Learning method that takes in an input image and gives priority (learnable weights and biases) to different characteristics and objects in the image to help it distinguish between distinct images. The output of the first convolution layer becomes the input for the second layer after the image has gone through one convolution layer. For each layer after that, this process is repeated. It is important to attach a completely connected layer following a series of convolutional, nonlinear, and pooling layers. The output data from convolutional networks is used in this layer. An N-dimensional vector, where N is the number of classes from which the model chooses the desired class, is produced by attaching a fully linked layer to the end of the network.

B. Recurrent Neural Networks (RNN)

The nodes in a directed graph which constitute a recurrent neural network (RNN) will be connected in the form of a series. It facilitates RNN's management of series operations like time sequence, handwriting recognition, and sequence expression. RNNs have the capacity to retain key details about the input they receive, enabling them to anticipate the subsequent item in a sequence. RNN usually features a STM and hence cannot handle very long sequences. Long STM (LSTM) network extends RNN which extends the memory of RNN. Therefore, LSTM are often employed in problems with sequences having long gaps. LSTMs can remember the previous inputs for an extended duration because it stores all those data into a memory. In image captioning problems, captions are generated from image features using RNN alongside LSTM.

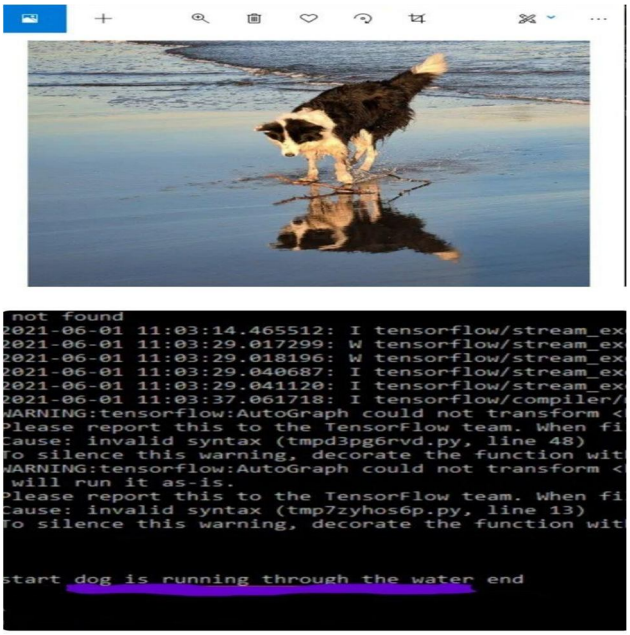
C. Long Short-Term Memory

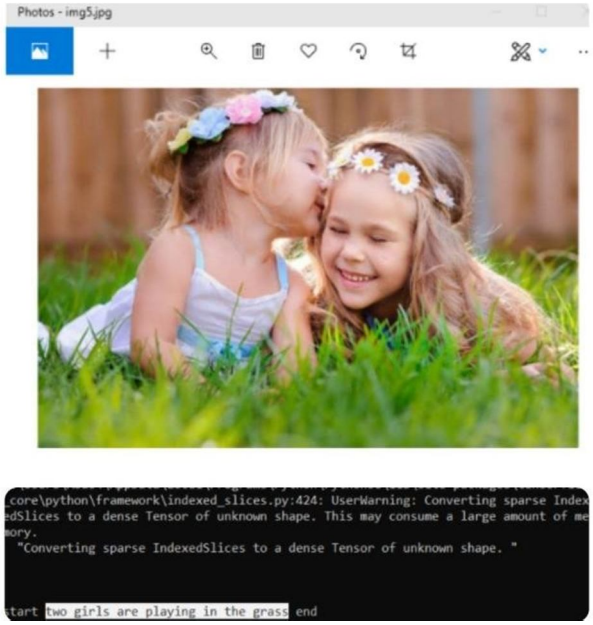
Recurrent neural networks (RNNs) of the Long Short-Term Memory (LSTM) type are able to recognize order dependence in sequence prediction issues. The most common applications of this are in difficult issues like speech recognition, machine translation, and other issues. When training conventional RNNs, this issue was observed because as we go further into a neural network, if the gradients are very small or zero, little to no training can occur, resulting in poor predicting performance. Since there may be lags of uncertain length between significant occurrences in a time series, LSTM networks are well-suited for categorizing, processing, and making predictions based on time series data. As it overcomes the short term memory constraints of the RNN, LSTM is significantly more efficient and superior to the regular RNN. The LSTM can process inputs while processing pertinent information, and it can ignore irrelevant information and performance.

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue V May 2024- Available at www.ijraset.com

VIII. RESULTS





IX. PROJECT CONCLUSION

In conclusion, our project on image caption generation using deep learning demonstrates significant strides in bridging the semantic gap between visual content and textual descriptions. Through meticulous model development, training, and evaluation, we've achieved notable success in generating contextually relevant captions for diverse images. Leveraging cutting-edge deep learning architectures and extensive datasets, our model exhibits proficiency in understanding complex visual contexts and producing linguistically coherent descriptions. The rigorous testing and validation procedures underscore its robustness and generalization capability, ensuring reliable performance in real-world scenarios. While our project marks a substantial milestone in the field of computer vision and natural language processing synergy, there are avenues for further enhancement, such as exploring multimodal approaches and refining caption quality. Ultimately, our endeavor contributes to advancing accessibility and comprehension of visual content, promising broader applications across industries ranging from assistive technologies to multimedia content enrichment.

©IJRASET: All Rights are Reserved

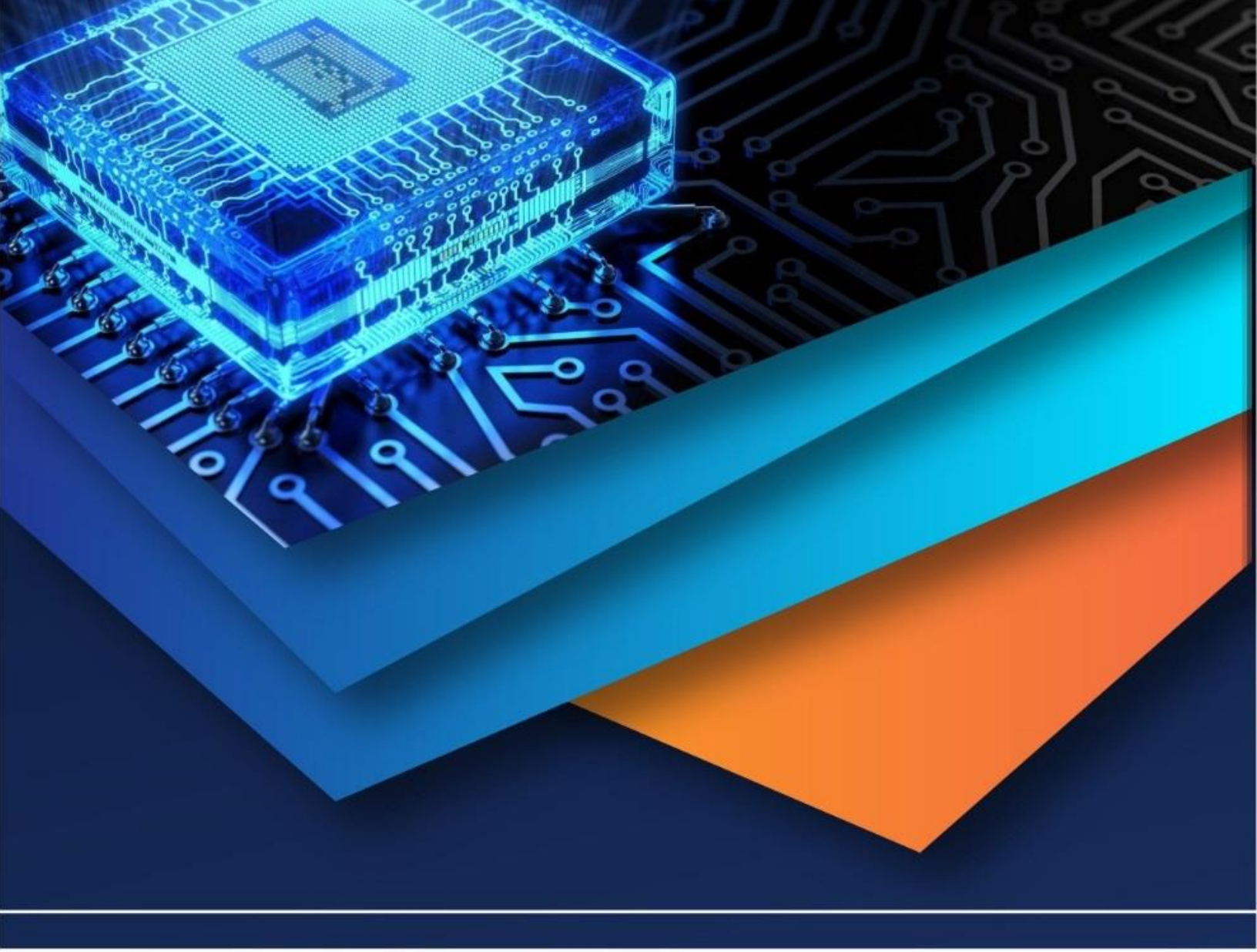
X. FUTURE ENHANCEMENT

The future scope for image caption generators using deep learning is promising. Further advancements could focus on enhancing model interpretability, enabling users to understand how captions are generated. Integrating multimodal inputs such as audio and text could enrich captioning capabilities, enabling more comprehensive understanding of visual content. Additionally, personalized captioning tailored to individual preferences and contexts could be explored, improving user engagement and satisfaction. Continued research into ethical considerations, including bias mitigation and privacy preservation, will be vital for responsible deployment. Overall, the evolution of image caption generators holds potential for revolutionizing communication, accessibility, and user interaction with visual media.

REFERENCES

[1] Marc‘ Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015.‖ Sequence level training with recurrent neural networks‖arXiv preprintarXiv:1511.06732.
[2] Peter Anderson, Xiaodong He, Chris Buehler, DamienTeney, Mark Johnson, Stephen Gould, and LeiZhang. 2017.‖ Bottom-up and topdown attention for image captioning and vqa‖. arXiv preprintarXiv:1707.07998.
[3] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. 2016‖. Selfcritical sequence training for image captioning.
[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, JamesHays, Pietro Perona, Deva Ramanan, Piotr Dollar, and C Lawrence Zitnick. 2014.‖ Microsoft coco:Common objects in context.‖ In European conference on computer vision, pages 740–755. Springer.
[5] Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan,and Eric P Xing. 2017

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



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)