JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Automatic Image Captioning With Natural Language Processing & Neural Networks

[1]Gayathri Devi.S , [2]Maheshwarlal.K, [3]Hari Prasanth.G

[1]Assosicate Professor, [2]Student-Final Year, [3]Student-Final Year
[1]Computer Science and Engineering,
[1]Sri Ramakrishna Institute of Technology, Coimbatore, India

Abstract : The fusion of natural language processing (NLP) and deep neural networks (DNNs) has propelled advancements in automatic image captioning, addressing the intricate challenge of bridging the semantic gap between visual content and descriptive language. This project investigates and implements a sophisticated framework that amalgamates NLP and DNNs to generate accurate and contextually rich captions for images. The initial phase of the project involves a comprehensive exploration of existing image captioning methodologies, emphasizing the limitations in understanding complex visual scenes and contextual nuances present in images. Leveraging the power of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the proposed system establishes a robust foundation for extracting high-level visual features and learning intricate linguistic patterns, respectively. The CNNs encode the salient visual features, while the RNNs decode these features into coherent and contextually appropriate textual descriptions. A pivotal element in the project involves pre-processing the image data to extract meaningful visual features through CNNs and leveraging pre-trained models. Simultaneously, the text data undergoes tokenization and sequence modeling to comprehend the sequential nature of language.

Keywords- CNN,NLP,RNN.

I INTRODUCTION

The fusion of Natural Language Processing (NLP) and Deep Neural Networks (DNN) has revolutionized the field of computer vision, particularly in the realm of automatic image captioning. This Module explores the seamless amalgamation of NLP and DNN, delving into their collaborative synergy in generating descriptive captions for images. Leveraging the power of deep learning and linguistic analysis, this study investigates the intricate process of teaching machines to not only comprehend the content of images but also articulate them in coherent, human-like language.

By exploring various models, methodologies, and the underlying technologies. This report aims to provide an insightful analysis of the challenges, advancements, and the potential future trajectory of automatic image captioning.

The integration of NLP and DNN stands as a pivotal milestone in the development of Artificial Intelligence (AI) systems capable of understanding and describing visual content autonomously. A novel approach to automatic image captioning that leverages the complementary strengths of NLP and DNNs. Our proposed method not only explores the semantic understanding of image content through deep neural networks but also incorporates linguistic constraints and syntactic coherence through natural language processing techniques.

Through the seamless integration of these two domains, we aim to achieve superior performance in generating accurate and human-like captions for a diverse range of images.

Traditional approaches to image captioning often relied on handcrafted features and rule-based systems, limiting their scalability and adaptability to diverse image datasets.

With the advent of deep learning techniques, especially convolutional neural networks (CNNs) for image processing and recurrent neural networks (RNNs) for sequential data modeling, the landscape of image captioning has undergone a paradigm shift.

The integration of these techniques with NLP methodologies has facilitated the creation of more robust and context-aware image captioning systems.

This causes a whole new trajectory in the field of the medical science and technology which is involved in the matriculated formation of Machine Learning Development forming the Major Leap forward. The Conclusive part is arranged in the form of an Ease App. formation of Machine Learning Development forming the Major Leap forward. The Conclusive part is arranged in the form of an Ease App.

On the following Based functions are majorly based on the technical field of medical related studies and either the formation of structural and medicological field is being followed here.The partial Visually impaired people and people who are partially impaired are so well benefited in this formation of system.

Abbreviations and Acronyms:

CNN- Convolutional Neural Network, LSTM- Long-Short Term Memory, NLP- Natural Language Processing, GRU- Gated Recurrent Unit, MS COCO- Microsoft Common Objects in Context, AI- Artificial Intelligence, VGG16- Visual Geometry Group 16, AMD- Advanced Micro Devices, TPU- Tensor Processing Unit.

II. METHODOLOGY

A hybrid approach combining natural language processing (NLP) techniques and deep neural networks (DNNs) is proposed to obtain automatic image registration .The approach consists of several key steps. First, a large data set containing images with corresponding subjects is collected.

This data set is important for training and evaluation of the proposed model. Common datasets such as MSCOCO or Flickr30k are considered. Second, a deep learning system is built to manipulate images and create captions. Convolutional neural networks (CNNs) are used to extract image features, capturing the visual information needed to create appropriate annotations.

At the same time, recurrent neural networks (RNNs) or transformer-based architecture are used to perform natural language generation, benefiting from sequential dependencies in titles. The proposed method provides a moving system beyond for automatic image captioning, which combines the capabilities of NLP and DNN to create more informative and accurate captions.

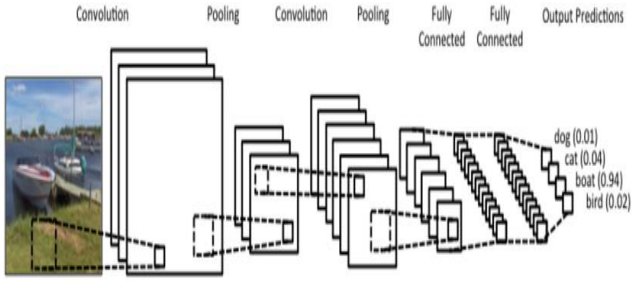

Fig. 1. Architecture of the system

By the Architecture and Fig. 1. We can clearly see the pooling layers in which the image gets segmented and therefore goes to fully connected layer (i.e): The pixelated image gets abbreviated into a form where the definitive each and every pixel is analyzed and studied according to the CNN Algorithm.

As per the aggregated time period the pixelated image goes to the epoch process and generates the Caption of the random or selected image with the accuracy and evaluation metrics, Also by interconnecting it to the webpage the following action is proceeded and then a final result is generated.

On the final format the Running engine is designed with CSS and HTML format as a webpage where through this connected ipynb file the mechanism is followed on the webpage content which is provided by a temporary IP Address for the following final result format.

III. ALGORITHM USED:

3.1 Convolutional neural networks (CNN):

It plays a crucial position in extracting functions from input photos. Convolutional Neural Networks (CNNs) are a class of deep neural networks commonly carried out to research visible imagery. In this context, CNNs are utilized to extract significant representations or features from photographs, which are then used as input to the captioning version. The CNN structure commonly includes more than one layers, which include convolutional layers, pooling layers, and fully linked layers. Convolutional layers follow filters to the enter photograph, extracting capabilities along with edges, textures, and patterns at exclusive spatial scales. Pooling layers down sample the function maps, lowering their spatial dimensions at the same time as preserving important records. Fully connected layers in addition method the extracted functions to generate a compact illustration that captures the image's content material.

In the assignment, the CNN set of rules is pre-trained on a large dataset (e.g., ImageNet) to research widespread-motive visible capabilities. These pre-trained CNN models, along with VGG, ResNet, or Inception, have tested sturdy overall performance in various pc vision duties and serve as function extractors for the photograph captioning machine. By leveraging transfer learning, the task blessings from the understanding acquired via the CNN version at some stage in pre-education, allowing it to effectively capture relevant image functions even if educated on a smaller dataset precise to the captioning venture. During the captioning method, the output of the CNN algorithm is usually fed right into a recurrent neural network (RNN) or transformer-based totally architecture accountable for generating textual descriptions. By combining the visible functions extracted by way of the CNN with

linguistic information from the captioning version, the device can produce accurate and contextually applicable photograph descriptions. This integration of CNNs with NLP strategies bureaucracy the inspiration of the project's approach to automated photograph captioning, enabling it to generate natural and coherent textual descriptions for a extensive range of input photographs.

3.2 Long-Short Term Memory (LSTM):

LSTM (Long Short-Term Memory) algorithm is a sort of recurrent neural community (RNN) architecture that is well-desirable for collection modeling duties. It addresses the vanishing gradient trouble that occurs in traditional RNNs by introducing specialised reminiscence cells and gating mechanisms. In the context of automatic image captioning, LSTM networks are hired to generate herbal language descriptions of snapshots based on their visible functions.

One key benefit of the usage of LSTM for photograph captioning is its ability to capture long-range dependencies in sequential information. This is specifically essential whilst producing captions for complicated photographs wherein contextual facts from extraordinary elements of the image can be important for generating correct descriptions. LSTM networks can effectively learn these dependencies over the years, allowing for more coherent and contextually applicable captions.

Additionally, LSTM networks are able to coping with variable-period input sequences, which is vital for picture captioning tasks where the length of the description can also range depending on the complexity of the image. By processing the visual functions extracted from the image through LSTM layers, the model can dynamically alter the period and shape of the generated captions to higher in shape the content material of the image. Furthermore, LSTM networks can be trained the usage of strategies which include trainer forcing, in which the model is fed the ideal target collection at on every occasion step throughout schooling. This facilitates to improve the stableness and convergence of the training manner, resulting in better overall performance on the captioning challenge. Overall, LSTM algorithms play a critical function in combining natural language processing with deep neural networks for computerized image captioning, allowing the era of descriptive and contextually relevant captions for a wide variety of images.

IV. SYSTEM ARCHITECTURE:

The various datasets of various datasets like random images presenting contents inside are randomly downloaded or placed in a packaged datasets called Flikcer8k, Flicker30k or MSCOCO are obtained and then passed on to the deliverance model.

The Deliverance model is either obtained from the Designed CNN and LSTM Model for the appliance of the CNN-LSTM Model the datasets are passed as training dataset as a part and also testing dataset which belongs to another variable format. As per the Dataset Classification the Mentioned amount of the datasets are either tested or ignored and finally the model studies the image.

During this process the epoch which is mentioned on the coding part is built in particular time period runs and during this period the model studies using the training dataset by based on the training dataset, The LSTM algorithm remembers the contents which was previously obtained on the training dataset and the final dataset is gathered by the CNN.

Therefore on the Final generation the Caption is catches by the LSTM and Generated including the image. As representing in a Flowchart format the architecture will be designed as mentioned below.

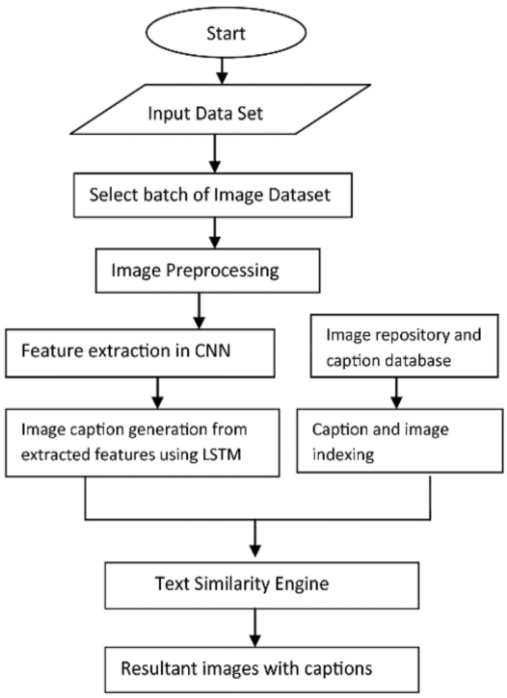

Fig.2. Working of the system

V. EXPERIMENTAL RESULTS:

Using the combined CNN-LSTM Module the Caption will generate us with the result by the evaluation metrics which includes graphical format.

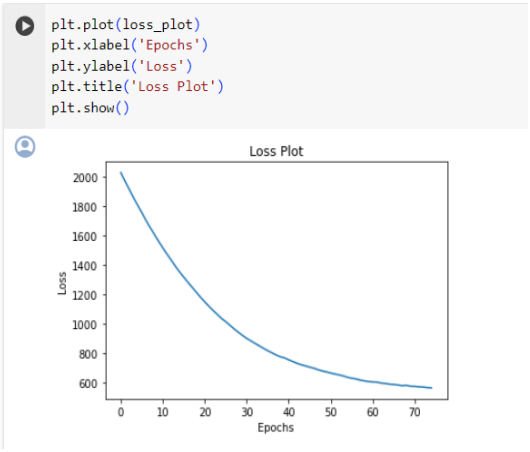

Fig.3. Output of Evaluation Metrics Using CNN-LSTM

By the Fig.3. Depicts the Loss part for the datasets which was a result of the loss of captions from the text dataset and by pixels which are blurred and hard to analyze. The x-axis indicated the Epochs which are completed and the Y axis indicates the loss of the Plotted Datasets.

As you can clearly see the in x-axis 1 cm = 10 units and in y-axis 1 cm = 200 units. The implementation was gathered on Tensorflow.

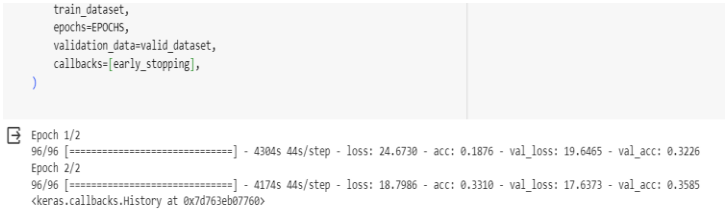

Fig. 4. Output of Epochs which are completed and on process

As we clearly see how much epochs are being generated here and what was the time taken for the epoch to run which relates with the loss function, accuracy, value_loss and value_accuracy.

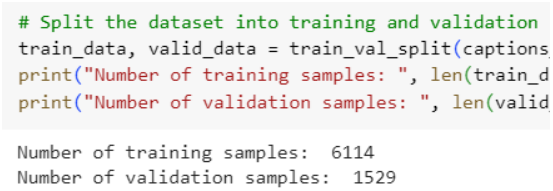

Fig. 5. Output of Classifying the Training Dataset and Testing Dataset

Splitting the Flicker_8k dataset into training dataset and testing dataset for the Generating and evaluating stages and later on comparing the Dataset according to the produced CNN Algorithm.

This Streamlit library gives the Webpage the overall attraction with inbuilt allocated variables which are helpful for developers.

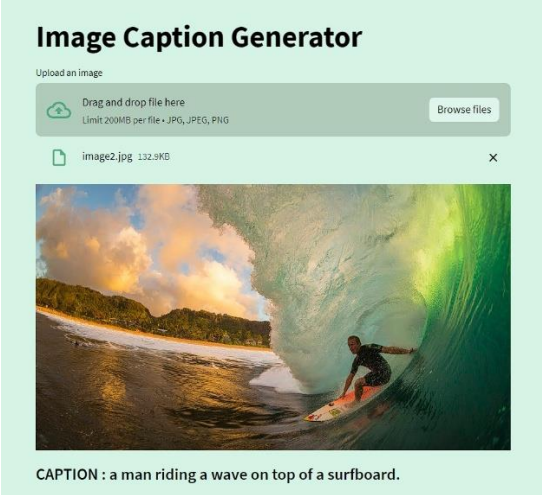

Fig. 6. Output of The Webpage which is connected with the CNN-LSTM Module

Streamlit is a popular open-source Python library used for building interactive web applications for data science and machine learning projects. It allows developers to create web-based interfaces quickly and easily, using simple Python scripts.

Therefore the Basic Webpage with connected User Interface to Generate the Captions with determined Buttons for browsing and generating the Captions.

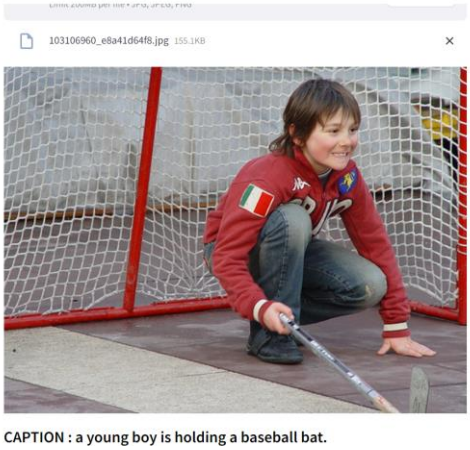

Fig. 7. Output of The Final Result obtained in webpage

With CSS, The webpage is enhanced and few additional features is being added on the Webpage which is converted as a Machine Learning Model.

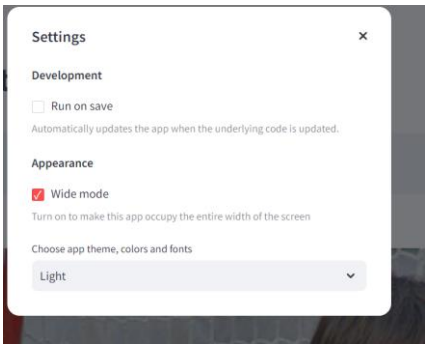

Fig. 8. Output of The Additional Features which are present on the Webpage

The Fig.8. Answers the module result generated using HTML, CSS and Streamlit, Where the Themes like Dark and White with including the System Default theme added. Also the Saving option is also generalized on the Webpage.

Even the Wider mode with desired size to view the panel and use the functions are also applied to the website and the button are designed and legible for the users to generate caption from any image.

VI. CONCLUSION:

This System of Module stands as an innovative endeavor that fuses the realms of computer vision and natural language processing to generate textual descriptions for images automatically. Through the integration of deep neural networks, particularly Convolutional Neural Networks (CNNs) LSTM and Transformer-based architectures, the project demonstrates a sophisticated approach towards understanding visual content and translating it into coherent, human-like captions. By leveraging the synergy between neural network models specialized in image feature extraction and language generation, the project excels in addressing the complex task of image understanding and caption generation. The application of Convolutional Neural Networks enables effective extraction of image features, while Transformer-based architectures handle the generation of textual descriptions by learning contextual relationships within the captions. Throughout the project, key methodologies have been explored and implemented, including the use of the Flickr8k dataset—a rich collection of images paired with multiple human-annotated textual descriptions. The dataset serves as a pivotal resource for training and evaluating the models, enabling diverse interpretations and captions for the same images, enriching the learning process and model performance.

Moreover, the adaptability of the CNN-LSTM framework allows for scalability and generalization across various domains and datasets, demonstrating its versatility and applicability in real-world scenarios. Future research directions may explore further optimizations and extensions of this hybrid architecture, such as incorporating attention mechanisms or exploring novel network architectures, to continue advancing the capabilities of automatic image captioning systems.

Overall, the successful integration of CNNs and LSTMs underscores the potential of combining deep learning techniques from computer vision and natural language processing domains, paving the way for more sophisticated and intelligent multimodal systems capable of understanding and describing visual content with human-like proficiency.

REFERENCES

[1] You, Q., Jin, H., Wang, Z., Fang, C. and Luo, J., 2016. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4651-4659).

[2] Rinaldi, Antonio M., Cristiano Russo, and Cristian Tommasino. "Automatic image captioning combining natural language processing and deep neural networks." *Results in Engineering* 18 (2023): 101107.

[3] Yao T, Pan Y, Li Y, Qiu Z, Mei T. Boosting image captioning with attributes. In Proceedings of the IEEE international conference on computer vision 2017 (pp. 4894-4902).

[4] Aneja J, Deshpande A, Schwing AG. Convolutional image captioning. InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 5561-5570).

[5] Lakshminarasimhan Srinivasan DS, Amutha AL. Image captioning-a deep learning approach. Int. J. Appl. Eng. Res. 2018;13(9):7239-42.

[6] Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V. Self-critical sequence training for image captioning. InProceedings of the IEEE conference on computer vision and pattern recognition 2017 (pp. 7008-7024).

[7] Feng Y, Ma L, Liu W, Luo J. Unsupervised image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2019 (pp. 4125-4134).

[8] Hirota Y, Nakashima Y, Garcia N. Quantifying societal bias amplification in image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022 (pp. 13450-13459).

[9] Mun J, Cho M, Han B. Text-guided attention model for image captioning. In Proceedings of the AAAI conference on artificial intelligence 2017 Feb 12 (Vol. 31, No. 1).