IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Developing And Instructing Models For Image Captioning Using Deep Learning Techniques

SHAIKH AZEEM AHMED SHAIKH WASEEM AHMED[1], Dr. V. S. KARWANDE[2]

ME Student, Department of Computer Science & Engineering, EESGOI , India[1]
HOD, Assistant Professor, Department of Computer Science and Engineering, EESGOI,India. [2]

Abstract: Picture subscription is a business that demands its consumers to have a strong understanding of both language and visuals in order to be successful. It is essential for image models to be able to create words in human languages, and this can only be accomplished through the interpretation of visual material. Image subtitle technology has made considerable use of the focus approach because it is able to deliver more in-depth sequential model training with more accurate picture information. This is one of the reasons why the technique has enjoyed such widespread adoption. The problem of automatically describing the contents of a photograph through the use of natural languages is one that is not only significant but also challenging to solve. This contains a variety of possible consequences that might occur. In order to illustrate this point, it can be useful in gaining an understanding of the fundamental characteristics of the picture. The skill of supplying picture information that is more exact and succinct is another one of its capabilities. This feature is important in scenarios such as the sharing of photographs on social networking sites. Deep neural networks are utilised in this research in order to achieve the goal of achieving this objective. Convolutional neural networks, often known as CNNs, are utilised in order to extract vectors from real-time video, which is represented by image frames. Following that, an LSTM network is employed in order to create replacements that are based on these vectors. In order to analyse the model, the dataset that is used the most commonly is the one that is taken from Flickr 8K. Approximately eight thousand photographs are included in it. The generation of picture captions is accomplished through this approach by gathering information from photographs and captions that are paired together. After that, these captions are utilised in the construction of photographs.

Keywords: Convolutional Neural Networks (CNN); Long Short-Term Memory networks (LSTM); Recurrent neural networks (RNN); Deep neural networks (DNNs).

I INTRODUCTION

The Jobs that are easy for people to accomplish but difficult for robots to execute include image captioning and object recognition, among many more. Nothing is wrong with deep neural networks (DNNs), which are very effective learning models that could accomplish remarkable results in challenging tasks such as object recognition, voice recognition, and image captioning. The title can only be constructed when the machine has grasped the scene and content of the image. The language model needs to be understood in order for the generated text to seem human and easy to interpret. Scientists have tried a number of approaches to train robots to accurately describe visual scenes, as humans struggle to do so. It takes more effort to caption photos than picture recognition since extra components and their relationships need to be

recognised, and then a concise title for the image needs to be written. Just a few examples of real-world applications include autonomous vehicles, navigation, image remote sensing, scene classification, and many more. Although deep learning has been around for a while, the increased availability of digital data and powerful GPUs have sped up recent advancements in the field. Thanks to the open source community, practical frameworks like PyTorch and Tensor Flow, large labelling datasets like Mscoco and Flicker, and excellent presentations, the field of deep learning is expanding at an exponential rate. Models for picture subtitling, including template, extraction-based, and encoder-decoder-based models, have been the subject of several research efforts in the past few years. Among these devices, the encoder decoder performs the best. This paper presents an architecture that incorporates a convolutional neural network (CNN), an encoder for visual information extraction, and a recurrent neural network (RNN) for phrase production.

Automated image subtitling has come a long way thanks to picture-derived semantic ideas. Our current ideas-to-caption method, on the other hand, is lacking in ideas since it trains the concept detector to reduce word disparities using picture-section pairings. For two main reasons: 1) the significant mismatch in the amount of positive and negative concept samples; and 2) poor marking in training titles as a result of the twofold annotation and the usage of synonyms. An approach to fixing the issues with online positive reminders and missing concepts (OPR-MCM) is discussed in this article. Using online positive retry predictions, our system reevaluates the loss of various samples and uses a two-stage optimisation technique for missing mining ideas. More semantic notions may be recognised by this method, and it has the potential to achieve high levels of accuracy. At each stage of the subtitling generating process, we use an element-by-element selection technique to find the most appropriate concepts. Because of this, our algorithm can generate an image title that is more precise and thorough. Our method outperforms rival techniques in picture subscription, as demonstrated by our extensive experiments with the MSCOCO online test server and the MSCOCO image subscription dataset. [1] [2][5].

II LITERATURE SURVEY

In this research, An ongoing difficulty in the field of computer vision and natural language processing is the synthesis of picture understanding with language comprehension. The last ten years have seen the rise of deep learning approaches as effective means of meeting this problem. In order to increase the quality of picture captions, several studies have concentrated on reducing dataset bias, using vision-language pre-training approaches, and creating better assessment tools. An outstanding feature is the usage of end-to-end models, which enable picture caption prediction without a pipeline of custom-built models or complicated data processing. The integration of deep learning, computer vision, and natural language processing has enabled considerable advancements in the field of automatic picture captioning, despite the field's many challenges. It is anticipated that future improvements in technology and deep learning models would lead to even more accurate caption creation. This study covers the evolution of picture captioning techniques through a thorough assessment of existing methodology and new advances. By zeroing in on important parts of the image, attention mechanisms greatly improve the relevancy and quality of produced captions. In addition, the article elucidates the difficulties of picture caption production, including diverse visual content and ambiguity in interpretation. In order to overcome these obstacles and improve captioning performance, many strategies are investigated, such as adversarial training and reinforcement learning. To wrap things off, this study delves into the latest advancements in picture caption creation and suggests avenues for further research. New opportunities for multimedia applications and human-machine interaction are emerging as a result of ongoing efforts to train computers to interpret and describe visual material through the integration of deep learning techniques with picture understanding and language production.[1]

It is possible to solve the difficulty of identifying the contents of a picture by utilising artificial intelligence. This may be accomplished through the utilisation of natural language processing (NLP) and computer vision (CV). The purpose of this research is to propose a generative model that

is used to produce genuine phrases that describe a picture. The model makes use of a deep recurrent architecture that integrates recent advancements in machine translation and CV. Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) are utilised in the development of an image caption generative model. This model employs several strategies of CV in order to comprehend the context of an image and describe it in a manner that is native to the English language. Specifically, the Flicker8K [17] dataset is utilised for the purpose of training. There are 8000 different photos included in it. The process of generating captions is one of the fascinating subfields of artificial intelligence that faces a great numerous obstacles. Following the successful generation of the caption, it is then turned into an audio format. The result demonstrates that the method that was built is producing results that are encouraging with regard to the development of captions.[2]

Using the Flickr8K dataset, this research looks at how well a picture captioning model that combines VGG16 and LSTM architectures performs. Extensive testing and analysis revealed the strengths and weaknesses of the model in producing picture captions with descriptive text. The results provide direction for future developments in the area of picture captioning techniques and add to the overall knowledge of these methods. Data preparation, model training, and assessment were all part of the investigation into VGG16 and LSTM architecture. A dataset called Flickr8K was used as a basis; it contains 8,000 photos with narrative descriptions. We trained LSTMs after preprocessing the data and extracting features with VGG16. To get the best possible performance, the model's parameters and hyperparameters were optimised. We used ROUGE scores, Semantic Similarity scores, and BLEU scores as our evaluation measures. The model showed a high level of semantic similarity, even if the BLEU score indicated modest overlap with reference captions. Analysis of ROUGE scores, however, showed difficulties in preserving coherence and capturing higher-order language patterns. The findings of this study have far-reaching consequences for fields including computer vision, NLP, and HCI. Improved accessibility, better picture interpretation, and easier human-machine connection are all possible outcomes of image captioning models that fill in the semantic gaps between visual material and written descriptions. There is need for development in terms of model design, attention mechanism integration, and the utilisation of larger datasets, even though the performance in collecting semantic content is promising. As the field of picture captioning continues to innovate, we may look forward to more sophisticated systems that will have many uses in many other fields.[3]

Over the past several years, there has been a growing interest in natural language processing and computer vision research on the issue of automatically producing descriptive phrases for photographs. Picture captioning refers to the practice of writing a textual explanation for a photograph or other visual medium. Computer Vision and Natural Language Processing are utilised in the production of the captions. A hybrid system that use a multi-layer Convolutional Neural Network (CNN) to generate image-descriptive vocabulary and makes use of a Long Short-Term Memory (LSTM) to accurately construct coherent sentences by making use of the keywords that were generated in this research. The term "Convolutional Neural Network" (CNN) refers to an approach for Deep Learning that use convolutional neural networks as its neural network architecture. Flickr8k, which has 8 thousand photographs, Flickr30k, which contains thirty thousand images, MS COCO, which contains one hundred eighty thousand images, and many more open-source datasets are available for use in relation to this subject.[4]

An ICG built using the VGG16 architecture for deep learning. Developing a model that can automatically generate contextually relevant and meaningful captions for input images is the primary goal of this research. To bridge the semantic gap between visual material and written descriptions, our technique makes use of VGG16's rich feature representations learnt from its pre-trained convolutional layers. In this proposed model, a VGG16-based convolutional neural network (CNN) is used by the encoder to extract picture features, while recurrent neural networks (RNNs) are used by the decoder to produce captions. The encoder uses VGG16 to effectively extract high-level features from pictures, which allows for efficient visual information encoding.

Captions generated by the decoder may then be tailored to the exact subject matter and surrounding environment of the input photos. The decoder's attention approaches improve the model's ability to create diverse and informative captions by focussing on relevant picture areas while each caption word is being formed. To further promote diversity and coherence in the produced captions, techniques such as beam search and vocabulary enrichment are employed. The proposed technique is demonstrated to generate informative captions for a range of photographs through test results on reference datasets such as Flickr8k. The model's ability to create captions that are both linguistically and contextually appropriate has been demonstrated through both qualitative and quantitative evaluations. This highlights the model's potential for several applications, including picture interpretation, retrieval, and accessibility for those with visual impairments.[5]

In recent years, thanks to breakthroughs in deep learning algorithms, the process of creating informative image captions has seen remarkable development. Producing coherent, contextually appropriate captions while fully comprehending visual material remains a formidable issue, even after considerable progress. We provide a new multimodal approach to picture captioning in this research that combines three strong deep learning architectures: Transformers for effective sequence modelling, EfficientNetB7 for efficient feature extraction, and YOLOv8 for robust object recognition. All three of these models—YOLOv8 for object detection, EfficientNetB7 for feature representation, and Transformers for contextual comprehension and sequence generation—are brought together in our suggested model. We demonstrate our approach's success in generating relevant and semantically rich captions for varied photos by conducting extensive tests on common benchmark datasets. The current state of the art in picture captioning jobs may be advanced by combining YOLOv8, EfficientNetB7, and Transformers, as demonstrated by the experimental findings. Captions for a wide variety of photos have been produced using the suggested multimodal technique, and they are both useful and rich in semantic information. The model has accomplished state-of-the-art performance in image captioning challenges by integrating the strengths of YOLOv8, EfficientNetB7, and Transformers. This method is significant because it tackles the difficult problem of interpreting visual material in its entirety while also producing cohesive and contextually appropriate captions. By combining three robust deep learning architectures, we can see how multimodal fusion may improve picture captioning by working together. Moreover, this method has far-reaching consequences for the industry, allowing for the development of more advanced and contextually aware picture captioning systems and launching fresh lines of inquiry into multimodal deep learning. Computer vision, natural language processing, and human-computer interaction are just a few of the areas that can benefit greatly from these systems.[6]

Captions are necessary for a wide range of functions in today's society, including the creation of news headlines based on images, the sharing of photographs on social networking platforms, and many more possible applications.

The goal of an image captioning system is to produce captions for each image automatically, as opposed to manually producing descriptions for each image themselves. In addition to providing a descriptive statement for a picture, it also helps in the semantic interpretation of the visual. The VGG16 algorithm is capable of using picture understanding, which is an essential way for decoding semantic image data. When it comes to the process of picture captioning, natural language processing and computer vision are both merged. In order to accomplish the objective, one may choose to implement either a conventional machine learning technique or a deep learning strategy. Identifying the objects and figuring out the connections between them are both necessary steps in the process of carrying out the project that has been planned. In order to turn the picture into a vector that can then be processed further, a technique known as feature extraction is utilised. Following the receipt of the items and the visual material, the LSTM will link the words in order to produce a sentence that provides a description of the things. The purpose of this study is to show the implementation strategy for Object Detection-based Picture Captioning that makes use of Deep Learning. When we were conducting the evaluation, we made use of the CIDEr metrics; the accuracy of 94.8% was reached for a total of thirty epochs, which is an extremely satisfactory result.

Image categorisation is accomplished through the use of CNN and the Flickr dataset.[7]

Image captioning is a very new and challenging subject that has brought a lot of attention to itself in recent times. At the moment, the goal of providing a concise description of an image in plain language is accomplished via the use of technologies that integrate computer vision (CV), natural language processing (NLP), and machine learning approaches. Within this body of work, we provided a description of a model that generates a description of a picture using natural language. The extraction of features was accomplished using the use of convolutional neural networks, and recurrent neural networks were utilised in order to create text based on these collected data. When we were developing captions, we made sure to take into account the attention mechanism. A model evaluation was carried out with the use of the Flicker8k database. Both encouraging and competitive are the results of the competition.[8]

The purpose of this project is to construct an image caption generator by using the power of deep learning. More specifically, we utilise Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks to do this. With this forward-thinking endeavour, the massive datasets and computing capabilities that are accessible in the field of deep learning are utilised to their full potential. The development of a system that is capable of understanding the content of an image and articulating it in English is our primary objective. This survey study introduces the core ideas of picture captioning and provides an overview of the many approaches that are often used. The Keras library, numpy, and Jupyter notebooks are some of the most important tools that we have at our disposal. In addition, we investigate the possibility of employing CNNs and the Flickr dataset for the purpose of picture categorisation within the framework of our research.[9]

A method for the production of picture captions that is based on deep neural networks is subjected to a comprehensive analysis in this article.The purpose of picture captioning is to bring about the generation of a sentence description for an image automatically. The image will serve as the input for our article model, which will then produce an English phrase as the output, which will describe the characteristics of the image. It has garnered a significant amount of interest from researchers in the field of cognitive computing in recent years. Due to the fact that the concepts of computer vision and natural language processing are mixed together, the work is considered to be rather complicated. Through the use of the principles of a Convolutional Neural Network (CNN) and a long Short-Term Memory (LSTM) model, we have produced a model. Additionally, we have constructed a functional model of an Image caption generator by using both CNN and LSTM methodologies. BLEU Scores are utilised in order to assess the effectiveness of our model following the completion of the caption creation phase. As a result, our system assists the user in obtaining a description that is descriptive for the image that is provided as input.[10]

In the field of artificial intelligence, one of the most fascinating and difficult tasks is to attempt to automatically describe the contents of a picture. An improved image captioning model that includes object identification, color analysis, and image captioning is suggested in this study. The purpose of this model is to automatically provide written descriptions of pictures. The VGG16 neural network is utilized as the encoder in an encoder–decoder model for the purpose of image captioning. The LSTM (long short-term memory) network with attention is utilized as the decoder in this model. In addition, the Mask R-CNN with OpenCV algorithm is utilized for the purpose of analyzing colors and detecting objects. After that, the process of integrating the picture caption and color recognition is carried out in order to give improved descriptive features and information about the photos.In addition to that, the sentence that was formed from the text is turned into speech. Furthermore, the results of the validation demonstrate that the suggested technique is capable of providing a more accurate description of pictures.[11]

The technique of creating explanations about what is happening in a picture is accomplished through the process of image captioning. Using the assistance of Image Captioning, descriptions are constructed that provide information on the pictures. Image captioning is a technique that has a wide range of applications, including the analysis

of vast numbers of unlabeled photos, the discovery of hidden patterns for machine learning applications, the construction of software that assists blind people, and the guidance of self-driving automobiles. Deep Learning Models are able to be utilized in order to accomplish this Image Captioning. As a result of the development of deep learning and natural language processing, it is now much simpler to produce captions for the photos that are provided. The process of picture captioning will be carried out with the assistance of neural networks in this study. The Convolution Neural Network (ResNet) is utilized as the encoder, which allows for the access to the image features. On the other hand, the Recurrent Neural Network (Long Short Term Memory) is utilized as the decoder, which is responsible for the generation of captions for the pictures with the assistance of the image features and the vocabulary that is constructed.[12]

III. SYSTEMS ARCHITECTURE

According to the system model, the architecture that has been presented includes It is a difficult undertaking that requires understanding of both verbal and visual language in order to properly caption images. It is necessary for image models to have an understanding of the subject matter of the pictures that they receive in order for them to be able to generate sentences in natural languages. It is common practice for image captioning projects to make advantage of the attention approach, which has the ability to supply more accurate picture data for the purpose of deeper sequential model training. The difficulty of automating the process of picture characterisation through the use of natural languages is both basic and complex. There has a great deal of promise. An example of its potential utility would be the interpretation of the content of images. Additionally, it has the potential to provide more accurate and concise picture information in circumstances such as the sharing of photographs on social media websites. This research takes use of deep neural networks in order to arrive at its conclusion. Using a convolutional neural network (CNN), feature vectors are extracted from photo frames that are taken in real-time. Subsequently, LSTM networks are utilised in order to create subtitles based on these feature vectors. The Flickr 8K dataset, which is used to test the model, is one of the most prominent datasets for picture captioning. It comprises more than 8,000 photographs and is used to evaluate the model. The technique creates image titles that are generally understandable and linguistically acceptable. This is accomplished by evaluating picture pairings and subtitles.

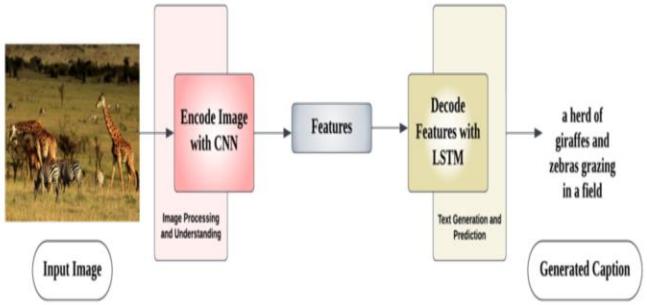

Figure No 3.1: System Architecture

IV EXPERIMENTAL RESULTS

During the process of developing images, one of the most typical problems that might occur is overfitting. When it comes to the complexity and ideal diversity of the captions that are made, there are very few training cases available. This is the reason why this is the case. We begin by addressing this issue by doing intensive hyperparameter optimisation on the dropout parameters. This is accomplished in the beginning. There is a lot of information that is shown in the graphic that is really interesting. The plot of measurements by epochs that is positioned on the right side of the graph reveals, first and foremost, that the majority of metrics at about epoch 5 are growing, and the CIDEr score is also increasing. This is demonstrated by the direction in which the graph is orientated.

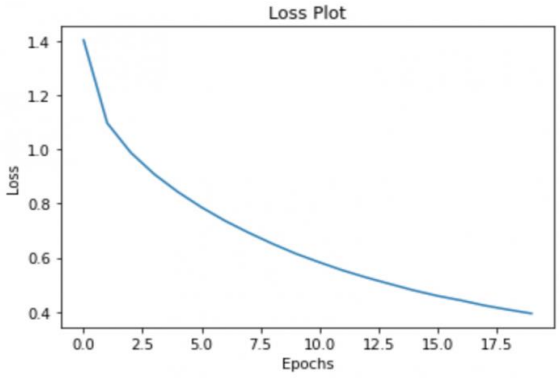

Figure No 4.1: Loss Function Curve.

V CONCLUSION

The purpose of this research is Visual Aligning Attention (VAA) and Deep Matrix Factorisation (DMF) are going to be combined in this study in order to produce a single model that will be used for the captioning of pictures. The purpose of this paradigm is to offer a solution to the problem of not teaching attention layers in a way that is understandable to students. CNN encoders are in charge of gathering visual information, whereas LSTM decoders are in charge of synthesising sentences that characterise the many elements that are contained inside images. The trained attention layers have the capacity to concentrate more accurately on regions and provide the decoder with more precise and helpful visual information. This is the most important aspect of the trained attention layers. Consequently, this makes it possible to generate phrases that provide a description of the contents of the photographs that are being input.

REFERENCES

1. Naresh Sharma,Hari Om,Chozharanjan,"From Pixels To Text: Deep Learning Approach For Image Caption Generation",IJNRD, Volume 9, Issue 3, March 2024.
2. Mr.Siddharaj D. Pujari and Mr. Atul S. Patane ," Image Caption Generation using Deep Learning Techniques",GIJET,2024.
3. Shan-E-Fatima,Kratika Gupta,Deepti Goyal,Suman Kumar Mishra,"Image Caption Generation Using Deep Learning Algorithm",Educational Administration: Theory and Practice(EATP),2024.
4. S. Suneetha, A.Meghana Laxmi Priya,Ch.Lakshmi Narayana, Ch. Enosh,T. Srinivasa Vara Prasad,"Image Captioning Using Deep Learning",IJERT,Volume 13, Issue 02, February 2024.
5. B. Bhaskar Rao,Kalki Chaitanya Lade,Avinash Pasumarthy,Parasuram Swamy Katreddy,Garapati Chaitanya Nagendra Kumar,"Image Caption Generator using Deep Learning Algorithm – VGG16 and LSTM",Volume: 11 Issue: 03,Mar 2024.
6. Rihem Farkh,Ghislain Oudinet1 and Yasser Foued,"Image Captioning Using Multimodal Deep Learning Approach",CMC,VOL.18,19 December 2024.
7. S.T.Santhanalakshmi,Dr. Rashmita Khilar,"Image Captioning Using Deep Learning",Journal of Harbin Engineering University,ISSN: 1006-7043,Vol 44, No. 7,July 2023.
8. Mr.N. Raghu,Sai Srikar,Aaftaab,Ruthvik Sai,"Image Captioning Using Deep Learning",IJRTI,Volume 8,Issue 4,ISSN: 2456-3315,2023.
9. T.Sandhya, Dr.Kondapalli Venkata Ramana ,"Generating Image Captions Based On Deep Neural Networks",IJNRD,Volume 8, Issue 9,September 2023.
10. Dr. S. Pasupathy,"Image Caption Generator by using CNN and LSTM",IJFMR,Volume 5, Issue 2, March-April 2023.
11. Yeong-Hwa Chang,Yen-Jen Chen,Ren-Hung Huang and Yi-Ting Yu,"Enhanced Image Captioning with Color Recognition Using Deep Learning Methods",MDPI,Appl. Sci. 2022.
12. Aishwarya Maroju,Sneha Sri Doma,Lahari Chandarlapati,"Image Caption Generating Deep Learning Model ",IJERT,Vol. 10 Issue 09, September-2021.
13. Yimin Zhou, Yiwei Sun, Vasant Honava,"Improving Image Captioning by Leveraging Knowledge Graphs",IEEE Conference,2019.
14. Soheyla Amirian,Khaled Rasheed,Thiab R. Taha,Hamid R. Arabnia,"Image Captioning with Generative Adversarial Network",CSCI,IEEE,2019.
15. Seung-Ho Han and Ho-Jin Choi,"Explainable Image Caption Generator Using Attention and Bayesian Inference",CSCI,IEEE,2018.
16. Vishwash Batra, Yulan He, George Vogiatzis ,"Neural Caption Generation for News Images",IEEE,2018.
17. Shiyang Yan,Fangyu Wu,Jeremy S. Smith,Wenjin Lu and Bailing Zhang,"Image Captioning using Adversarial Networks and Reinforcement Learning ",ICPR,August 20-24, 2018.
18. Lakshminarasimhan Srinivasan, Dinesh Sreekanthan, Amutha A.L,"Image Captioning - A Deep Learning Approach",IJAER,Volume 13, Number 9,2018.

19. Aghasi Poghosyan,Hakob Sarukhanyan,"Long Short-Term Memory with Read-only Unit in Neural Image Caption Generator",IEEE,2017.