IMAGE CAPTIONING USING DEEP LEARNING

[1]Mr.N. Raghu, [2]Sai Srikar, [3]Aaftaab , [4]Ruthvik Sai

[1]Assistant Professor, Department of Artificial Intelligence,
Anurag Group Of Institutions,
[2,3,4]Student,
[2,3,4] Department of Artificial Intelligence,
Anurag Group Of Institutions, Hyderabad, India,

***Abstract*— Image captioning is a new difficult problem that has recently attracted a lot of attention. Currently, technologies that combine computer vision (CV), natural language processing (NLP), and machine learning techniques are used to complete the task of producing a brief description of an image in natural language. We described a model that creates a natural language description of an image in this work. Convolutional neural networks were combined to extract features, and recurrent neural networks were utilized to generate text from these features. When creating captions, we took the attention mechanism into consideration. Using the Flicker8k database, we assessed the model. The outcomes are encouraging and competitive.**

Keywords: Image captioning, convolutional neural networks, recurrent neural networks, attention mechanism.

I. INTRODUCTION

Every day, we are bombarded with pictures from many sources such as the internet, news stories, document diagrams, and ads. These sites include visuals that visitors must interpret for themselves. The majority of photos do not contain an explanation, yet humans can interpret them without them. However, if people want automated picture captions from robots, they must be able to comprehend a wide range of image descriptions. Image captioning is critical for a variety of reasons. Captions for each image on the web can lead to faster and more detailed image searches and indexing.

This project's purpose is to produce relevant descriptions for a given image. Captions will be developed to collect contextual information from the photographs. To create acceptable captions, current approaches employ convolutional neural networks (CNNs) and recurrent neural networks (RNNs) or versions thereof. To accomplish this objective, these networks utilize an encoder-decoder technique, in which CNNs encode the picture into feature vectors and RNNs serve as decoders to provide linguistic descriptions.

Image captioning is used in a variety of industries, including commerce, biomedicine, online search, and military applications. Captions for photographs can be generated automatically by social media platforms such as Instagram and Facebook.

II. LITERATURE SURVEY

As an initial step, a comprehensive review of the existing literature was conducted.

[1] Xiangqing Shen, Bing Liu, Yong Zhou, and Jiaqi Zhao et al. describe remote sensing picture caption creation using transformer and reinforcement learning. They propose a novel model that uses the Transformer to interpret picture information into target phrases, and then use Reinforcement Learning to improve the quality of the created phrases. The proposed model is validated on three remote sensing image captioning datasets. It achieves the highest scores on the Sydney Dataset and Remote Sensing Image Caption Dataset (RSICD), as well as higher scores on the UCM dataset, outperforming earlier state-of-the-art models in remote sensing image caption production.

[2] Nayan Mehta, Suraj Pai, and Sanjay Singh explain automated 3D sign language caption creation for video. The purpose of this research is to provide a useful solution for leveraging internet resources and making them available to hearing-impaired people in their primary form of communication. The video is first translated to text using subtitles and speech processing technologies. NLP algorithms understand the produced text, which is then translated to avatar captions, which are then displayed to make a coherent movie alongside the original material. It is found that students taught with sign captioned films outperform those taught with English captioned videos by 37% and 70%, respectively, as verified across a 6-month period and a subsequent 2-month trial.

[3] Chetan Amritkar, Vaishali Jabade, and colleagues discuss image caption generation using deep learning technique. This approach is used to produce natural phrases that describe the image, utilizing Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). CNN is utilized for picture feature extraction, whereas RNN is employed for phrase production. The model is trained to create captions that roughly represent the image when a picture is fed into it. The accuracy of the model and the command of the language model learned from visual descriptions are examined on various datasets. The trials demonstrate that the model frequently provides appropriate descriptions for an input image.

[4] Philip Kinghorn, Li Zhang, Ling Shao, and their colleagues propose a novel deep learning architecture that generates image descriptions based on regions, with an emphasis on enriched and detailed descriptions. The proposed method involves using a regional object detector, attribute prediction through recurrent neural networks (RNNs), and an encoder-decoder language generator

that utilizes two RNNs. This approach is aimed at enhancing existing holistic techniques by focusing on local-based methods, especially for image regions that contain people and objects. The system was evaluated using the IAPR TC-12 dataset and achieved superior performance compared to state-of-the-art techniques across various evaluation metrics.

.

III. PROBLEM STATEMENT

The task of image captioning represents a formidable challenge that entails the generation of a natural language depiction of an image. Despite the ease with which humans are able to undertake this task, machines encounter considerable obstacles owing to the intricate and heterogeneous nature of natural language. The current era has witnessed considerable progress in the area of deep learning methodologies, with a specific focus on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These advancements have shown great potential in the realm of image captioning. Notwithstanding, the task of generating precise and semantically significant captions persists as a noteworthy hurdle. Furthermore, contemporary methodologies commonly exhibit deficiencies in their capacity to explicate objects and scenes with ample precision, thereby yielding insufficiency or even inaccuracy in the corresponding captions. Thus, the objective of the present study is to introduce a pioneering deep learning framework for image captioning, which is capable of producing comprehensive and precise descriptions through the exploitation of both overarching and specific characteristics of the image. The present study vows to subject the proposed model to a rigorous evaluation using prevalent benchmark datasets. Furthermore, the comparative analysis will be carried out in relation to contemporary cutting-edge techniques, with a view to assess the performance of the model.

3.1 Existing System

Recently, there has been a notable surge in interest towards image captioning, particularly in the domain of natural language processing. The exigency for contextually appropriate descriptions of images in natural language is of utmost importance. Contemporary developments in fields such as neural networks, computer vision, and natural language processing have cleared the way for accurate depiction of images and their visual-based significance, although this notion may initially appear implausible.

3.2 Proposed System

The proposed model is to generate the relevant natural language caption to the given input image, instead of just describing a single target object the model detects multiple target objects for generating grammatically correct caption

.

IV. IMAGE CAPTIONING METHODS

Various techniques for generating image captions exist, some of which are underutilized in contemporary times. However, it is imperative to comprehensively survey these technologies before proceeding further. The current strategy for generating subtitles for pictures can be delineated into three principal classifications.

The technique of image captioning that utilizes templates as a foundation for generating descriptive textual content is alternatively referred to as template-based image captioning in academic discourse.

The subject matter under consideration pertains to retrieval-based image captioning.The concept of generating novel captions is a burgeoning field of research in contemporary academia.

Many methodologies for image caption generation utilizing deep learning techniques are categorized as novel caption generation. Accordingly, our attention is directed to the exclusive area of newly devised caption creation through the use of deep learning. The image captioning methodology that primarily employs deep machine learning techniques and visual space is the novel caption generation-based technique. The classification of deep learning-based image captioning techniques can also be attributed to various learning methodologies.

The following list outlines several topics related to computer vision that are often discussed in academic writing. These include visual space-based approaches, multimodal space-based approaches, supervised and unsupervised learning techniques, dense captioning methods, whole scene-based approaches, and encoder-decoder architecture based models.

(8) The approach of Compositional Architecture-based captioning involves utilizing a carefully designed network architecture to generate image descriptions. (9) The LSTM (Long Short-Term Memory) language model-based technique employs a recurrent neural network structure that can learn from a sequential input sequence and generate captions for images. (10) The Attention-Based method is characterized by identifying and focusing on relevant image features while producing captions. (11) The Semantic concept-based approach leverages the understanding of concepts to describe the image contents. (12) The Novel object-based captioning technique endeavors to identify and describe newly encountered objects in an image through learning patterns and cues from previous objects.

4.1 Retrieval-based approaches

Captions can be obtained from both visual and multimodal space. Captions are collected from collections of existing captions in retrieval-based systems. Retrieval-based approaches begin by locating visually related photos with captions in the training data set. These are known as candidate captions. The captions for the query image are chosen from this pool of

captions.These approaches generate captions that are both broad and syntactically valid. They cannot, however, construct image-specific and semantically precise captions.

4.2 Novel-based approaches

Captions can be created from both visual and multimodal space. This category's main technique is to analyse the image's visual content and then construct image captions from the visual material using a language model. These algorithms can create semantically more accurate captions for every image than earlier efforts. Deep learning-based approaches are used in the majority of innovative caption creation systems. As a result, unique picture captioning algorithms based on deep learning are our primary emphasis.

V. DEEP LEARNING BASED IMAGE CAPTIONING METHODS

5.1 Visual Space vs. Multimodal Space

Image captioning systems based on deep learning may create captions from both visual and multimodal space. Image captioning databases, understandably, include text captions. The picture characteristics and accompanying captions are individually provided to the language decoder in the visual space-based approaches. In contrast, a shared multimodal space is learnt from the pictures and caption-text in a multimodal environment. The language decoder is then given this multimodal representation.

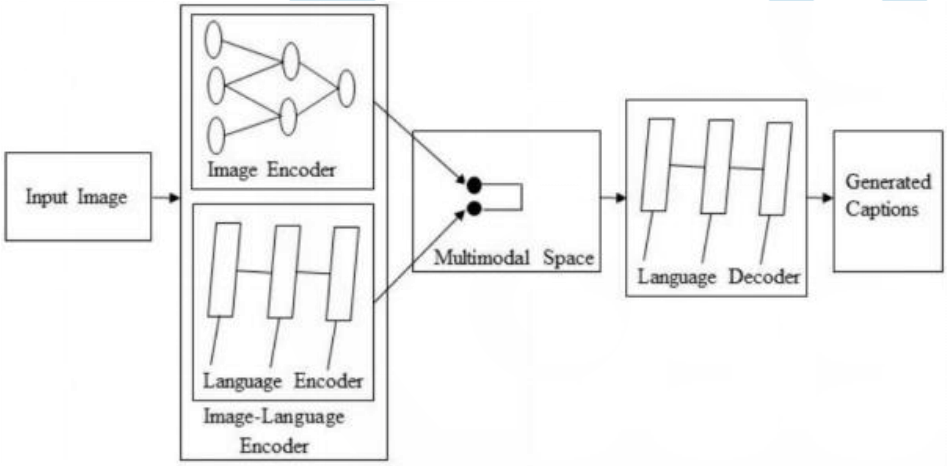

Figure 5.1: A block diagram of multimodal space-based image captioning.

5.2 Multimodal Space

A ubiquitous strategy that employs diverse modes of space-based communication entails the integration of several key components, including a linguistic Encoder, a visual component, a multimodal spatial element, and a Linguistic Decoder. Figure 2.1 depicts a schematic representation of the methodologies for space-based multi-modal image captioning. Within this framework, the visual component employs a deep convolutional neural network to function as a feature extractor and isolate relevant image features. The language encoder component facilitates the extraction of salient linguistic traits and subsequently assimilates a compact and cohesive characteristic representation for each lexeme. Subsequently, the semantic temporal context is transmitted to the recurrent layers. The process of mapping image features into a unified space with word features is performed by the multimodal space component. The map that ensues is subsequently transmitted to the language decoder, through which captions are foretold via map decoding.

The techniques encompassed within this particular classification adhere to a prescribed sequence of actions:

In the context of a multimodal environment, the combined use of deep neural networks and multimodal neural language models is applied for the purpose of conducting joint learning of both textual and visual representations within a unified multimodal space.The segment dedicated to language production produces captions by utilizing the data obtained from the preliminary step.

5.3 Supervised Learning vs. Other Deep Learning

In supervised learning, training data is accompanied by the expected output, which is referred to as labels. Unsupervised learning, on the other hand, works with data that has not been labelled. Unsupervised learning techniques such as Generative Adversarial Networks (GANs) are used. Reinforcement learning is another type of machine learning technique in which an agent's aim is to find data and/or labels through exploration and a reward signal. A variety of photo captioning methods employ reinforcement learning and GAN-based algorithms.

5.4 Encoder-Decoder Architecture-Based Image captioning

End-to-end picture captioning systems based on neural networks operate in a basic manner. Based on the encoder-decoder paradigm, these approaches are quite similar to neural machine translation. In this network, global picture properties are extracted from CNN's hidden activations and input into an LSTM to create a series of words. This category's usual procedure includes the following main steps:

1.A vanilla flavour CNN is used to determine the scene type, as well as to recognise objects and their connections.

2.A language model is utilised to turn the output of Step 1 into words, combined sentences, and image captions.

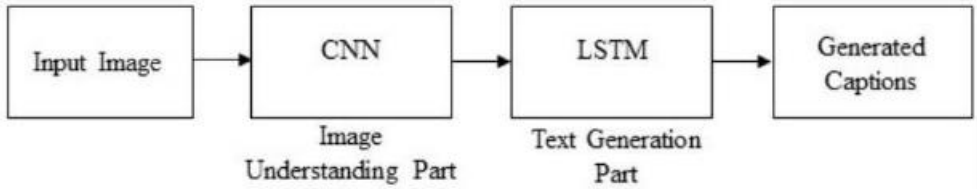

Figure 5.1: A block diagram of simple Encoder-Decoder architecture-based image captioning

VI. IMPLEMENTATION

This project's major purpose is to learn about deep learning techniques. We generally use two approaches for picture captioning: CNN and LSTM. As a consequence, we'll mix these designs to develop our model of picture caption generator. The CNN-RNN model is another name for it. To extract features from a picture, the CNN algorithm is utilised. The VGG16 model, which has already been trained, will be used.

The CNN input will be used by LSTM to assist construct an image description.

6.1 CONVOLUTIONAL NEURAL NETWORK

A convolutional neural network (ConvNet/CNN) constitutes a deep learning mechanism that is designed to receive an input image and allocate significance through learning weights and biases to different elements or objects within the image while simultaneously facilitating discernment among them. In comparison to alternative classification methods, the utilization of a ConvNet necessitates a notably reduced degree of pre-processing.

Convolutional Neural Networks (CNNs) are a distinct category of deep neural networks which possess the ability to process information inputted to them in the form of either a 3D or 2D matrix. Efficient representation of images can be achieved in a 2D or 3D matrix format, wherein the employment of Convolutional Neural Networks (CNNs) is highly advantageous for processing such data.

The process involves a sequential scanning of images from left to right and top to bottom to discern key features, which are subsequently integrated to facilitate their classification. The software possesses the ability to process images that have undergone a series of rotations, resizing, and subsequent rotations, in addition to modifications in perspective.

6.2 VGG16

The VGG16 represents a Convolutional Neural Network (CNN) architecture that emerged as the winner of the 2014 ILSVR (Imagenet) competition. This vision model architecture has gained recognition as a paradigm of exceptional quality. One notable characteristic of the VGG16 model is its emphasis on utilizing 3x3 filter convolution layers with a stride of 1, in lieu of a vast number of hyper-parameters. Additionally, the model consistently employs identical padding and a 2x2 filter stride 2 maxpool layer. In the architectural design, the convolution and maximum pooling layers are organized in a congruous manner. The object under consideration possesses a pair of FCs.

The ultimate layer comprises fully interconnected layers, culminating in a softmax function for output. Notwithstanding, the ultimate layer shall be eliminated as solely the image's features are requisite. The nomenclature "VGG16" denotes that the model comprises a total of sixteen stratified layers, each endowed with variant weight parameters. The magnitude of this network is formidable, as it encompasses an estimated 138 million parameters. The provided model was trained using the Imagenet dataset, which consists of 14 million photographs that are classified into 1000 distinct categories.

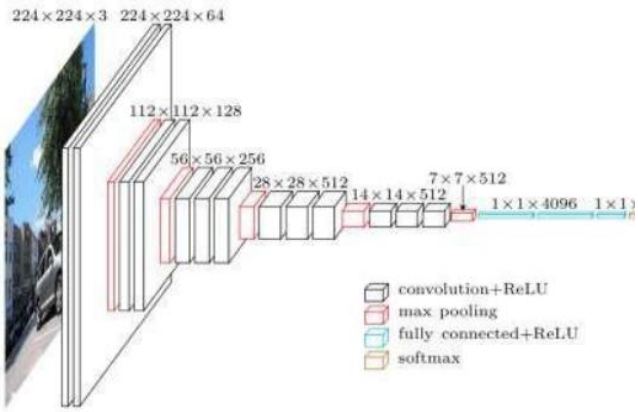

Figure 6.1 architecture of VGG16

6.3 LONG SHORT TERM MEMORY LSTM

Long short term memory (LSTM) is a kind of RNN (recurrent neural network) that excels at sequence prediction tasks. Based on the previous paragraph, we can estimate what the next word will be. In terms of overcoming the restrictions of RNNs with short term memory, it beat ordinary RNNs. The LSTM may carry out useful information throughout input processing and discard irrelevant information using a forget gate.

LSTMs, as opposed to traditional RNNs, are designed to avoid the problem of vanishing gradients and store information over long periods of time. LSTMs can maintain their error constant, allowing them to learn back propagation across time and layers over many time steps.

6.4 Process

Step1: First we need to install the Flickr8k dataset in the kaggle notebook and access GPU to increase the speed of processing.
Step2: The copied path of the caption data file and image file separately and will access the data one by one from disk.
Step3: VGG16 trained CNN models are used to extract the features of the images.
Step4:To prepare the text data, description of the images is to be cleaned. For this, all the words are converted to lowercase.
Step5:The text data is encoded and stored in a form of a directory.
Step6: A model has to be defined in order to display the process.It will be divided into two parts.
Sequence Processor – The textual input will be handled by an embedding layer.
Decoder - To make the final prediction, output from the above layers is combined.
Step7:BLEU score metric is used to evaluate the model.It requires comparing original caption with generated caption.
Step8:Required image is given to the model and a suitable caption is generated for the image.

VII. RESULTS

By using the combination of CNN and LSTM model we have generated an image captioning model. The model was trained using the Flickr8k dataset and after many epochs it is able to generate accurate captions to the images which are given to the model. The BLEU score, which is used to evaluate the model,is obtained as **0.54**.

```
---------------------Actual---------------------
startseq black dog and spotted dog are fighting endseq
startseq black dog and tri-colored dog playing with each other on the road endseq
startseq black dog and white dog with brown spots are staring at each other in the street endseq
startseq two dogs of different breeds looking at each other on the road endseq
startseq two dogs on pavement moving toward each other endseq
--------------------Predicted--------------------
startseq two dogs play with each other in the grass endseq
```

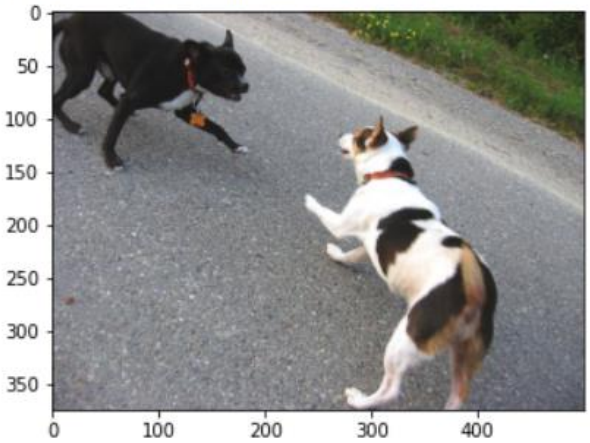

Fig 7.1 caption for image-1

```
---------------------Actual---------------------
startseq man in hat is displaying pictures next to skier in blue hat endseq
startseq man skis past another man displaying paintings in the snow endseq
startseq person wearing skis looking at framed pictures set up in the snow endseq
startseq skier looks at framed pictures in the snow next to trees endseq
startseq man on skis looking at artwork for sale in the snow endseq
--------------------Predicted--------------------
startseq two people are hiking up snowy mountain endseq
```

Fig 7.2 caption for image-2

VIII. CONCLUSION

We looked at deep learning-based picture captioning algorithms in this report. While deep learning-based picture captioning systems have made significant progress in recent years, a robust image captioning approach capable of producing high-quality captions for practically all photos has yet to be developed. Automatic image captioning will remain a hot research topic for a while, thanks to the emergence of novel deep learning network architectures. We used the Flickr 8k dataset, which contains over 8000 photos as well as the captions for each image. Despite the fact that deep learning-based picture captioning systems have made significant progress in recent years, a reliable image captioning method capable of producing high-quality captions for practically all photos has yet to be developed. Automatic image captioning will remain a hot research topic for quite sometime, thanks to the emergence of innovative deep learning network architectures. Because the number of users on social media is growing by the day, and the majority of them will share images, the scope of image captioning will be enormous in the future. As a result, this project will be of greater assistance to them.

REFERENCES

[1]Hao Zhou, Wei Huang, Zhuo Xiao, Shichuan Zhang, Wangzhan Li, Jinhui Hu, Tianxing Feng, Jing Wu, Pengcheng Zhu, Yanchao Mao, "Deep-Learning-Assisted Noncontact Gesture-Recognition System for Touchless Human-Machine Interfaces", Advanced Functional Materials, 2022,doi:10.1002/adfm.202208271.

[2]SaiTeja. N.R, Rashmitha Khilar, "Implementing Complexity in Automatic Image Caption Generator using Recurrent Neural Network over Long Short-Term Memory", Journal of Pharmaceutical Negative Results, Volume13,Special Issue , 2022.

[3] Peerzada Salman syeed, Dr.Mahmood Usman, "Image Caption Generator UsingDeep Learning", Neuroquantology, October 2022, Volume 20, Issue 12, Page 2682-2691, Doi: 10.14704/Nq.2022.20.12.Nq77261.

[4] SiZhen Li, Linlin Huang, "Context based Image Caption using Deep Learning", 2021 IEEE 6th International Conference on Intelligent Computing and Signal Processing (ICSP 2021), DOI: 10.1109/ICSP51882.2021.9408871.

[5] Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, And Pushpak Bhattacharyya, "A Hindi Image Caption Generation Framework Using Deep Learning", Trans. Asian Low-Resour. Lang. Inf. Process. 20, 2, Article 32 (March 2021), 19 pages, doi:10.1145/3432246.

[6] Xiangqing Shen, Bing Liu, Yong Zhou & Jiaqi Zhao, "Remote sensing image caption generation via transformer and reinforcement learning", Multimedia Tools and Applications, Fig. 5: Image Caption Generation Vol 12 Issue 02, Feb 2023 ISSN 2456 – 5083 Page 579 volume 79, pages 26661–26682 (2020), doi: 10.1007/s11042-020-09294-7.

[7] Nayan Mehta, Suraj Pai,Sanjay Singh, "Automated 3D sign language caption generation for video",Universal Access in the Information Society volume 19, pages 725–738 (2020).

[8] Chetan Amritkar, Vaishali Jabade, "Image Caption Generation using Deep Learning Technique",2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018 IEEE, 978-1-5386- 5257-2/18.

[9]Philip Kinghorn, Li Zhang, Ling Shao, "A region-based image caption generator with refined descriptions",Neurocomputing, Volume 272, 10 January 2018, Pages 416-424, Elsevier, Doi:10.1016/j.neucom.2017.07.014.

[10] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, Yoshua Bengio, "Neural Image Caption Generation with Visual Attention", Proceedings of the 32 nd International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.