Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Performance Enhancement of Image Captioning Technique Using Machine Learning Approach

Deepti Goyal[1] Dr. S. V. A. V. Prasad[2]
[1]Research Scholar, Lingaya's Vidyapeeth, Faridabad
[2]Dean (R&D), Lingaya's Vidyapeeth, Faridabad

Article Info
Page Number: 2054-2066
Publication Issue:
Vol. 71 No. 3s2 (2022)

Article History
Article Received: 28 April 2022
Revised: 15 May 2022
Accepted: 20 June 2022
Publication: 21 July 2022

Abstract
The present research works aims to attain highest accuracy to increase the performance of model used for image captioning. Henceforth, the in this proposed research an attempt was made using VGG 16 and LSTM algorithm. The fusion of computer vision and natural language processing has received a lot of interest recently thanks to the advent of deep learning. This field is represented by picture captioning, which trains a computer to comprehend an image's visual information using one or more phrases. The process of constructing a coherent description of high-level image semantics requires an understanding of the state, attributes, and relationships of these objects. Although picture captioning is a difficult and involved task, several researchers have made significant progress in this area. In this research, we focus on employing convolutional neural networks (CNNs) as encoders and recurrent neural networks (RNNs) as decoders to propose methods for photo captioning In the present report, image captioning is handled using the VGG16 and LSTM algorithms. Compared to other models, this one achieves the greatest accuracy (96.75%) and the lowest loss (0.22) of those that are currently available. Hence, the proposed model performed outstandingly.

Keywords: Image Processing, CNN, RNN, VGG16, LSTM, Highest Accuracy, Minimal Loss Function.

1. INTRODUCTION

Computer vision has advanced significantly in the area of image processing during the last few years, including image classification and object detection. The issue of image captioning, which involves automatically generating one or more phrases to comprehend an image's visual information, has benefited from advancements in image categorization and object detection. News picture titles, medical image descriptions, text-based image retrieval, information available to visually impaired users, and human-robot interaction are just few of the potential domains where automated creation of full and natural image descriptions might have a significant influence. Both theoretical and practical, these captioning-related uses are very relevant to the field. Because of the rise of AI, captioning images has become a more difficult yet crucial task. When presented with a new picture, a captioning system should provide a meaningful description of it. The sentence at the image's base explains the picture's content, the items that are slowly making their way into it, the action, and the setting.

Captioning photographs is a complex work that requires the integration of image processing, computer vision, natural language processing, and other important fields of study. However, this activity is readily accomplished by humans. The difficulty of image captioning is in

developing a model that can fully use picture data to provide more detailed, human-like descriptions of images. In order to provide meaningful description using high level picture semantics, it is necessary to be able to assess the states of the objects or scenes in the image, comprehend their relationships, and produce a phrase that is both semantically and syntactically valid. It is not yet known how the brain processes visual information to form a caption.

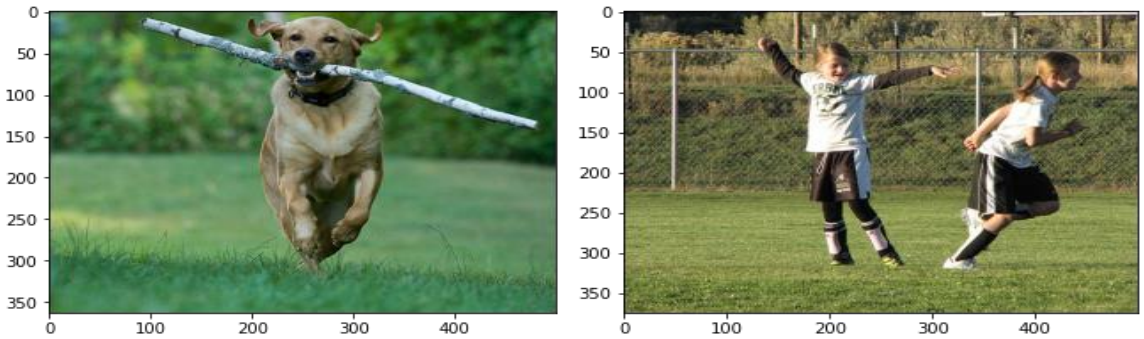

Fig no-1 Image being processed for caption

1.2. CNN AS ENCODER AND RNN AS DECODER

A picture is made up of several colors in a human's eyes to represent various scenes. Most visuals displayed on a computer screen, however, are formed using pixels in three different color channels. On the other hand, all of the neural network's data modalities are converging on the same destination: the generation of a vector and the subsequent application of certain operations to its features. CNNs have been shown to be effective in a variety of vision tasks, including object identification, detection, and segmentation, by providing a rich representation of the picture through embedding into a fixed-length vector. For this reason, CNN is a popular choice for the encoding step of encoder-decoder-based picture captioning methods.

The RNN network learns from the past via its hidden layer, which provides superior training capabilities and outperforms extracting deeper linguistic knowledge such as the semantics and syntactic information latent in the word sequence. If you look at the hidden layer state of a recurrent neural network, you can easily define the dependency connection between different location terms in the past. Images are captioned using an encoder-decoder technique, with the encoder being a convolutional neural network (CNN) model for feature extraction. ResNet, GoogleNet, VGG [10], and AlexNet are just some of the models it can employ. For decoding, the framework provides the word vector expression to an RNN model. The word embedding model takes each word's initial representation as a one-hot vector and expands it to the same dimension as the picture feature. The encoder and decoder in this suggested study were VGG16 and LSTM, which are algorithms from CNN and RNN, respectively.

1.3 CONVOLUTIONAL NEURAL NETWORK (CNN)

For problems with computer vision, CNN is a common neural network design. In order to locate relevant features, CNN has the advantage of automatically extracting functions from snapshots. During the flattening process, the image's geographic locations are erased. In order to preserve the spatial link between picture components, internal feature representation is trained using tiny squares of input data.

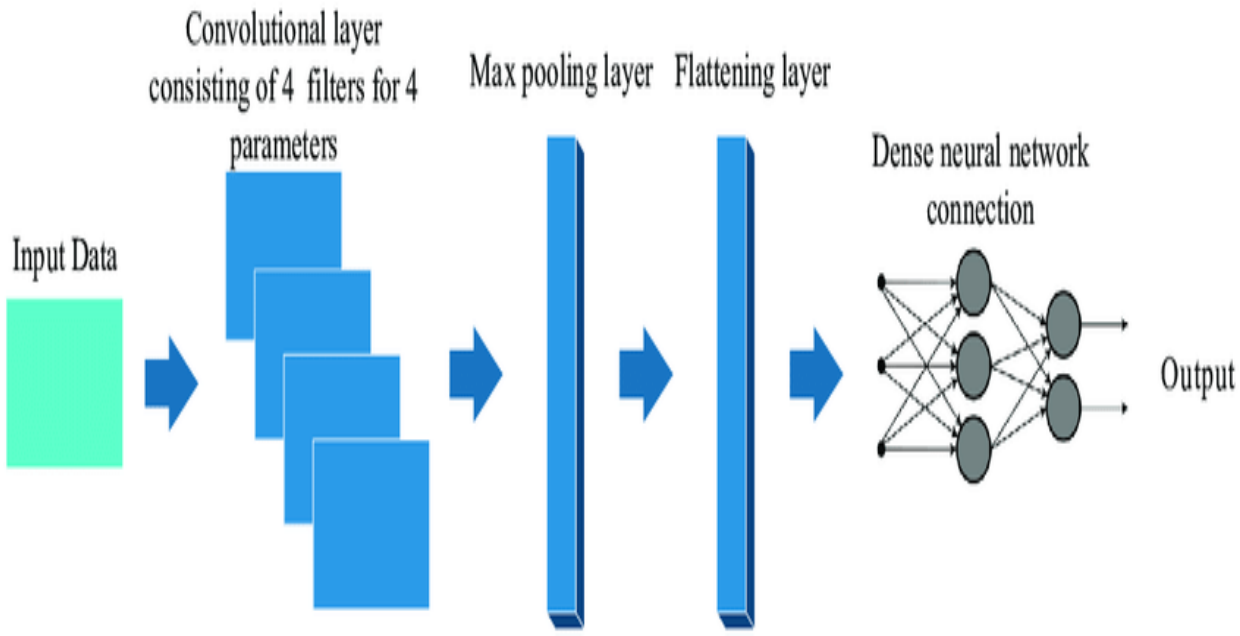

Fig no-2 an overview of CNN architecture

a) Convolutional Layer: Filters and feature maps make up the convolutional layer. Consciousness on a certain layer is processed by what is known as filters. There is no way to replace those filters. They begin with the values of the pixels and end up with a distinctive map. The outcome of one clean-out layer is a feature map. Clear out is applied to the whole image, one pixel at a time. A limited number of diverse neurons are triggered to provide a feature map. **(b) Pooling layer:** This layer is utilized to reduce dimensionality. Pooling layers are added after one or two convolutional layers to generalize the advancements discovered from prior feature maps. By doing this, the risk of over fitting throughout the academic year is decreased. **(c) Fully linked layer:** After gathering and combining features from the convolutional layer and pooling them afterward, the fully linked layer is used to allocate the feature to change. Linear activation capabilities or softmax activation characteristics are used in these layers.

1.3.1 VGG 16

The VGG16 convolutional neural network (CNN) architecture was used to triumph in the 2014 ILSVR (ImageNet) contest. The architecture is considered to be cutting-edge in the field of vision models. The most notable aspect of VGG16 is that it consistently uses the same padding and maxpool layer of 2x2 filters with a stride 2 and gives more priority to having convolution layers of 3x3 filters with a stride 1 than previous versions. The convolution and

max pool layers are laid out in the same way over the whole design. An image with dimensions is sent to the network (224, 224, 3). The first two tiers both use the same 33 sized filter and have 64 channels of padding. After that, a max pool layer with stride (2, 2) lies on top, and then two layers of convolution with 128-by-128-pixel filters (3, 3). The next layer is also a max-pooling stride (2, 2) layer, therefore it has the same characteristics. Then, there are 256 filters, split between two convolution layers of size 3 and 3. Following a max pool layer, there are two sets of three convolution layers. To ensure uniformity, all 512 filters have the same padding (3, 3). This image is then sent on to a stack of two convolution layers.

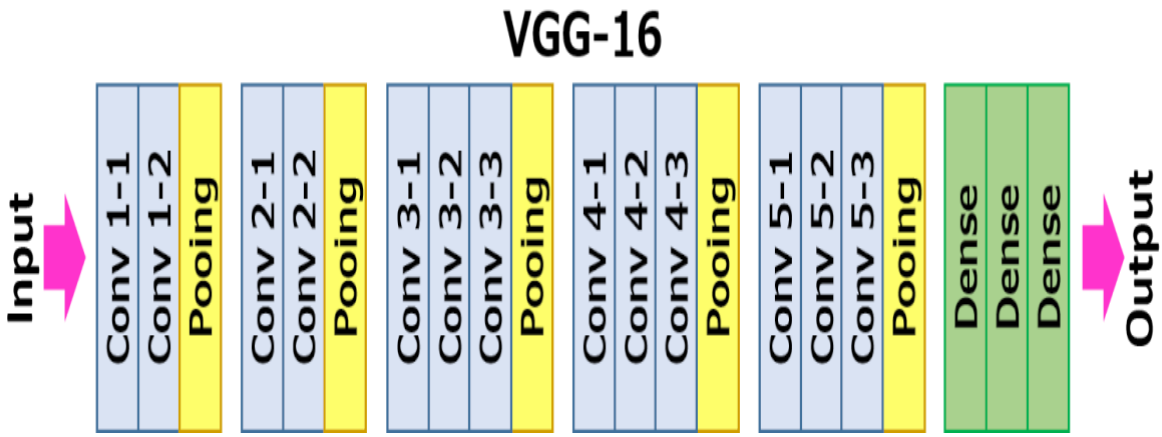

Fig no-3 an overview of VGG Model

1.4 RNN- LONG SHORT TERM MEMORY (LSTM)

LSTM networks are an extension of RNNs; they were first designed to deal with RNN failure cases. There are no changes made to the training model, and the vanishing gradient issue is virtually resolved. Because of its ability to handle noise, dispersed representations, and continuous input, LSTMs are often employed to bridge lengthy time lags in complex problems. As opposed to the hidden Markov model, LSTMs need not need to remember a fixed number of states in advance (HMM). Since this is the case, exact adjustments are unnecessary. The gating unit or gated cell in the LSTM's buried layer is the key structural difference between the LSTM and RNN architectures. It consists of four layers that work together to generate the cell's state and output.

These two components are passed on to the next concealed layer after that. When compared to RNNs, which only include a single tanh layer, LSTMs are more complex networks consisting of three logistic sigmoid gates and one tanh layer. Using gates, we are able to limit the quantity of data that may be sent through the cell. They decide what data the next cell will require and what data may be safely disregarded. The typical result is between 0 and 1, where 0 signifies "reject all" and 1 means "include all."

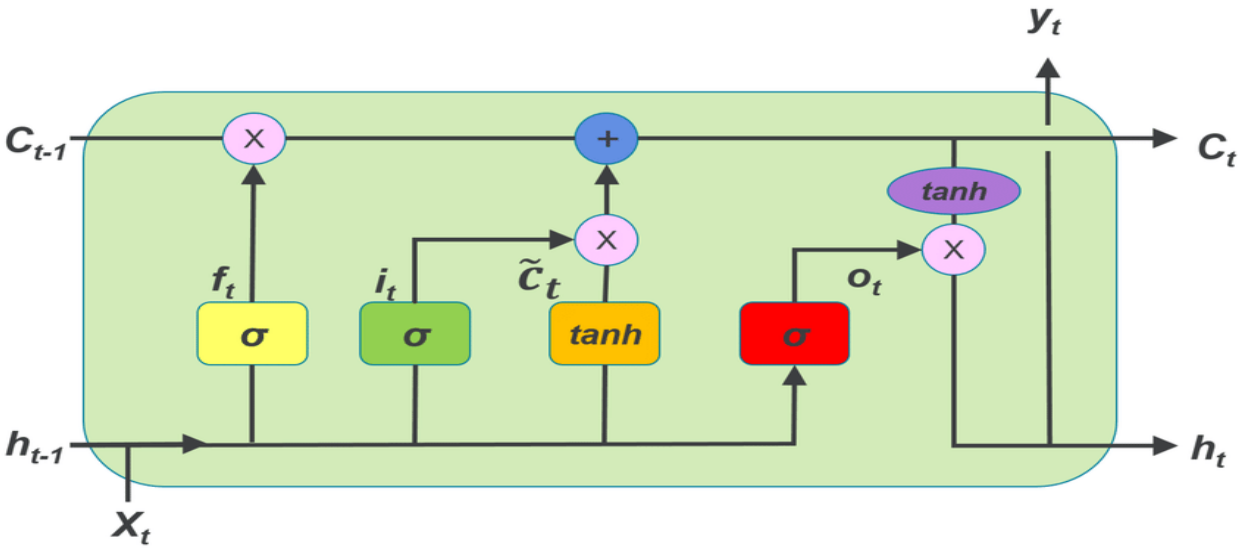

Fig no-4 an overview of LSTM Model

2. LITERATURE SURVEY

A user-friendly machine learning system for captioning images was created. It is a method for an interactive picture captioning learning model. It focuses on three primary areas: model updating, data augmentation, and feedback collection. It mostly uses the Top Down Approach and Standard encoder-decoder captioning methods. Consequently, it enhances picture captioning. Its accuracy rate is good, coming in at roughly 95.46 percent **(Mareike Hartmann et al., 2022)**. A development of a novel for modeling of Hyper parameter tuned Deep learning for automated Image Captioning. It encompasses two major parts namely Encoder and Decoder. It had been carried out against two benchmark datasets. Hence, it employs the faster Squeeze Net with RMS Prop model for extraction of visual features that exist in image . It gives the accuracy percentage of about 93.47% (**Mohamed and Sayed., 2022).** An attempt was made to provided a machine learning and image compression-based implementation and optimization of an image processing algorithm. It uses Tensorflow, Keras, and Python to caption images. The reduction of data size and execution time is the key goal here. The algorithm employed in this has an accuracy of roughly 91.67 percent and is based on CNN, PCA, and SVD (**Georgias Zacharis et al., 2022)**. The authors have devised a method for area recognition in ultrasound images using captioning. This method provides annotation text data to explain the illness contents in an ultrasound image while simultaneously encoding and detecting the focus region in ultrasound images, using the LSTM to decode the vector with accuracy rate of betTheyen 70-80%. The transformer then produces semantic and spatially explicit text descriptions. The sentence quality is then improved using the Reinforcement Learning (RL) method **(Zeng et al., 2022)**. A research study from the year 2020 proposes a technique for picture captioning utilizing multimodal features fusion with a mask. It makes use of a method of image captioning that enables computers to analyze a photograph's content and generate textual descriptions of its subjects. Deep learning is a major foundational component. Then, the python framework is used to build it, and performance measures like PSNR and RMSE are used to assess how well it does its job of interpreting images, resulting in an accuracy of 75-82.43%. (**Kumaravel**

Thangavel et al., 2022). These researchers have presented a new approach to picture captioning by fusing many types of information utilizing mask recurrent neural networks and LSM. Mask Recurrent Neural Networks (Faster R-CNN) are utilized in the coding layer, while long short-term memory (LSTM)-attend is used in the decoding process, and an accuracy of 70.48 percent is achieved in the creation of the descriptive text. Python frameworks are used for the paper's implementation, and several performance measures (including PSNR, RMSE, SSIM, Accuracy, Recall, F1-score, and Precision) are analyzed and discussed. **(Suresh Muthusammy.et.al., 2022)**.

Neural Attention for Image Captioning: A Review of State-of-the-Art Techniques is a proposed study. They do not provide an exhaustive analysis of all previous work on deep image captioning models but rather discuss the different kinds of attention processes utilized for the job in deep learning models. The encoder-decoder architecture is employed by the most effective deep learning models for picture captioning. Modern attentive deep learning models for picture captioning often utilize soft attention, bottom-up attention, and multi-head attention to achieve an accuracy of 75.66 percent. (**Jugal K Kalita.et.al., 2022).** Improving Image Captioning by Guessing Gaze Patterns from the Caption is the subject of recent study. In order to automatically generate a series of looked-at items given a fresh collection of photos, they first train a pointer network to learn from the captions how to predict the sequence of the stared objects. They test the effectiveness of the proposed sequence by pointer network in combination with preexisting picture caption models. (**Rehab Ahmed and James Hahn., 2022)**. The authors submit a study on Deep Learning for Image-Based Storytelling. Use of digital picture object recognition is a key component. They automatically produce a textual travelogue by combining the shifts in the spatiotemporal domain with the completion of a predefined template. Contrasted with traditional picture captioning, our goal in this project is to more efficiently link correlation digital images with contextual details. To do this, they employed deep learning in conjunction with object detection. (**Yulin Zhu and Theyi Qi Yan., 2022).** A Hybridized Deep Learning Approach to Bengali Image Captioning is offered by the authors. With the sole publicly accessible Bengali dataset for image captioning, BanglaLekha, they provide a standard technique for Bengali picture caption creation on two distinct sizes of the Flickr8k dataset. They also used a hybrid method that combined a Convolution Neural Network (InceptionResnetV2 or Xception) with a Long Short-Term Memory (Bidirectional Gated Recurrent Unit) and two Bengali datasets. Finally, Bilingual Evaluation Understudy was used to measure how well the suggested model worked, and it was shown to be superior on the Bengali dataset of 4,000 pictures and the BanglaLekha dataset, with an accuracy rate of 81.97 percent. (**Mayeesha Humaira.et.al., 2021)**.

A research article titled Towards Accurate Text-based Image Captioning with Content Diversity Exploration has been developed by the authors. They suggest a unique Anchor-Captioner approach. In particular, they look for the tokens that are meant to be given greater weight and use them as anchors. Finally, they use multi-view caption creation based on several ACGs to increase the resulting captions' subject variety. In addition to achieving SOTA performance, experimental findings demonstrate that our system also creates a wide

range of descriptive captions for photos. (**Mingkui Tan.et.al., 2021)**. An exhaustive survey of deep learning techniques for semantic segmentation and image captioning has been compiled by the authors. Using a fully convolutional network and other high-level hierarchical feature extraction methodologies, they describe the application of deep learning techniques to the segmentation analysis of 2D and 3D images, with an accuracy rate of about 81%. In the last part, we analyze the current methodologies, their contributions, and their applicability, emphasizing the importance of these methods and illuminating a possible research area for the use of semantic picture segmentation and image captioning techniques. **(Ariyo Oluwasammi.et.al., 2021).** To learn about and assess racial biases in image captioning, a model was presented. In particular, they focus on the COCO dataset to examine bias transmission processes in picture captioning. While previous studies have examined gender bias in captions via the use of machine generated gender labels, the present study instead examines racial and intersectional biases through the use of hand annotated photos. Furthermore, they demonstrate that these variations are more pronounced in contemporary captioning systems than in earlier ones, raising worries that, absent careful analysis and mitigation, these differences will only get worse and throws the accuracy rate of about 90.79% (**Dora Zhao.et.al., 2021)**. A deep learning-based picture captioning system for automatically generating extensive descriptions of bridge damage is presented in this work. Furthermore, it is very uncommon for bridge photographs to show more than one kind of damage; as a result, our system is customized to generate numerous words, allowing for a more complete understanding of intricate images. Scores range from 0.782 to 0.749 to 0.711 to 0.693 on the Bilingual Evaluation Understudy (BLEU) scale in our sample, with 69.67% of explanation phrases being properly generated. (**Yu Maemura(2021).**

A model was proposed a research paper on Image Captioning Through Self-Supervised Learning. They explored two solutions for the image captioning using two different self-supervised learned models, based on Jigsaw Puzzle solving and SimCLR, as a pre-text task. For the sake of supervised and self-supervised pre-text tasks comparison, They provide the results of their comprehensive testing on the same downstream task, calculating a BLEU score and validation loss. Our proposed solution with SimCLR model used for image feature extraction achieved the following results: BLEU-1: 0.575, BLEU-2: 0.360, BLEU-3: 0.266, BLEU-4: 0.145, accuracy rate of 74% and validation loss of 3.415. These outcomes can be considered as competitive ones with the fully supervised solutions (**Abdelrahman Mohamed., 2020)**.

3. RESEARCH METHODOLOGY

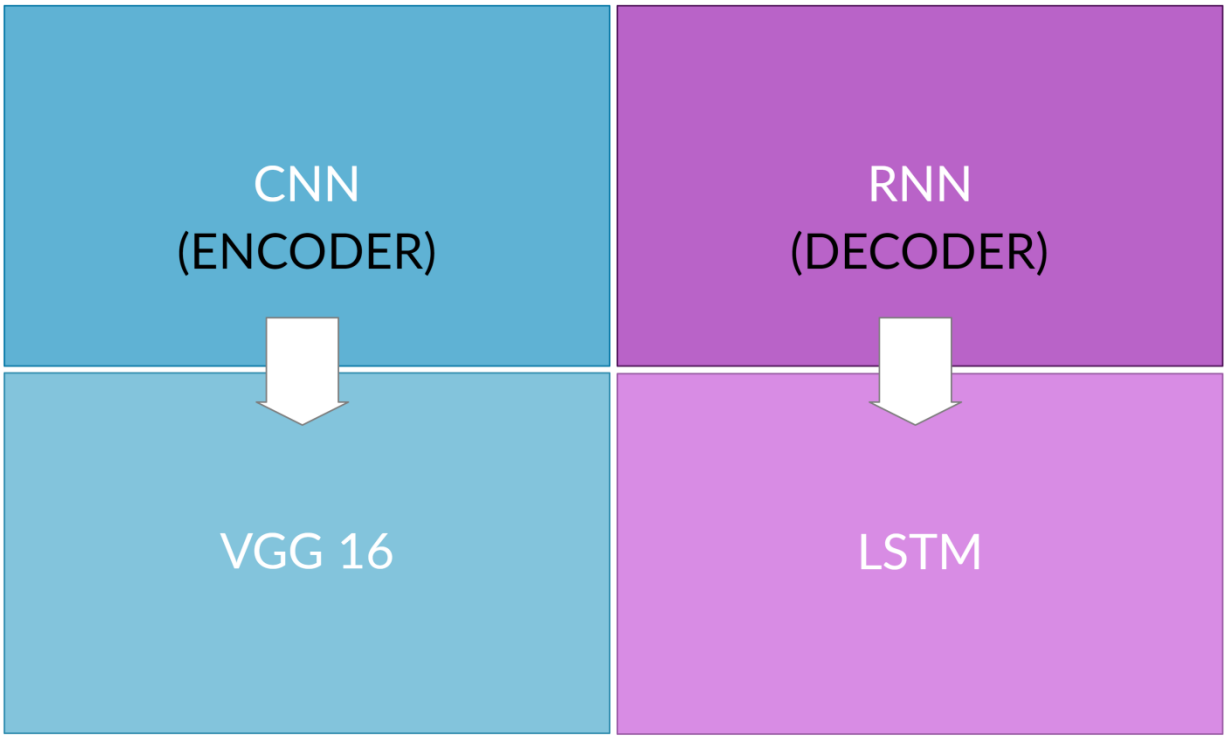

Fig no-5 The proposed model development approach

3.1 DEVELOPMENT PHASE

I. **First of all, the necessary software packages were installed.**

II. **Data Cleaning**

a) **load_doc (filename)** – In order to open the file and convert its contents into a string.

b) **all_img_capt (filename)** – Map photos using all five descriptions by creating a description dictionary.

c) **cleaning_txt (descriptions)** – In order to sanitize the data, this procedure accepts all available descriptions. When working with textual data, it is necessary to execute numerous sorts of cleaning, such as converting uppercase to lowercase, removing punctuation, and eliminating words containing numbers.

d) **text_vocabulary (descriptions)** – The unique terms that are culled from these descriptions are utilized to build a lexicon.

e) **save_descriptions** (descriptions, filename) – This feature creates a file containing all the descriptions before processing them.

III. **Extraction of features of the dataset**

IV. **Loading of a dataset to train the model**

The following functions were require performing to train the datasets:

a) **load_photos (name)** – This method accepts a filename as an argument and returns a string containing the file's contents, which is a list of picture filenames.

b) **load_clean_descriptions (name, image)** – This method creates a dictionary containing all of the captions for the images in the list.

c) **load_features(photos)** – This method would provide a dictionary for images as well as feature vectors collected from the suggested model.

V. **Tokenizing the vocabulary of the dataset**

VI. **Creation of a Data generator**

a) **Feature Extractor** – The feature is extracted from the size photos with the help of a dense layer, and the dimensions are reduced to less than zero nodes.

b) **Sequence Processor** – The LSTM layer would follow it. This embedded layer would handle the textual input.

c) **Decoder –** The final forecast will be made by combining the results from the two layers above and processing the thick layer.

VII. **Training the Image Caption Generator model is done**

VIII. **Finally, the testing of the Image Caption Generator model will be completed.**

3.2 MODEL TRAINING AND TESTING USING GOOGLE COLAB

```
Vocabulary Size: 7577
Description Length:  32
Model: "model_2"
__________________________________________________________________________________________________
 Layer (type)                   Output Shape         Param #     Connected to
==================================================================================================
 input_7 (InputLayer)           [(None, 32)]         0           []

 input_6 (InputLayer)           [(None, 512)]        0           []

 embedding_2 (Embedding)        (None, 32, 256)      1939712     ['input_7[0][0]']

 dropout_4 (Dropout)            (None, 512)          0           ['input_6[0][0]']

 dropout_5 (Dropout)            (None, 32, 256)      0           ['embedding_2[0][0]']

 dense_6 (Dense)                (None, 256)          131328      ['dropout_4[0][0]']

 lstm_2 (LSTM)                  (None, 256)          525312      ['dropout_5[0][0]']

 add_2 (Add)                    (None, 256)          0           ['dense_6[0][0]',
                                                                  'lstm_2[0][0]']

 dense_7 (Dense)                (None, 256)          65792       ['add_2[0][0]']

 dense_8 (Dense)                (None, 7577)         1947289     ['dense_7[0][0]']

==================================================================================================
Total params: 4,609,433
```

Fig no-6 Training of model using Google Colaboratory

4. RESULT AND DISCUSSION

This section presents the outcome of the trained model tested on the dataset which has colour images and the trained model generated the caption for the following images.

Start dog runs through the grass end **start man is standing on the grass end**

<matplotlib.image.AxesImage at 0x7f616b440610> <matplotlib.image.AxesImage at 0x7f6254782a10>

Fig (a) A trained dataset with caption “Dog running on the grass”

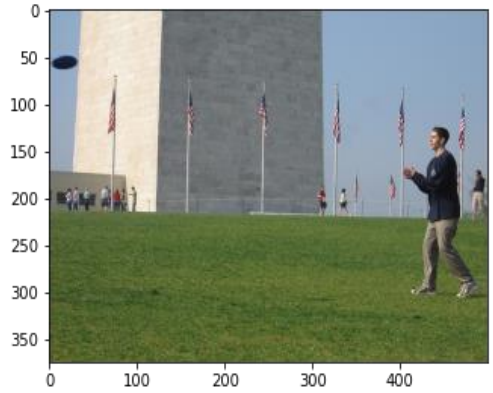

Fig (b) A trained dataset with caption “A man in black t shirt”

start man is standing on the sidewalk end start two girls are playing in the grass end

<matplotlib.image.AxesImage at 0x7f61675a94d0> <matplotlib.image.AxesImage at 0x7f616b472690>

Fig (c) A trained dataset with caption “man standing on the side wall”

Fig (d) A trained dataset with caption “Two girls are playing in the grass”

start man is standing on the edge of mountain end start dog runs through the grass end

<matplotlib.image.AxesImage at 0x7f6167a92d50> <matplotlib.image.AxesImage at 0x7f62546ba850>

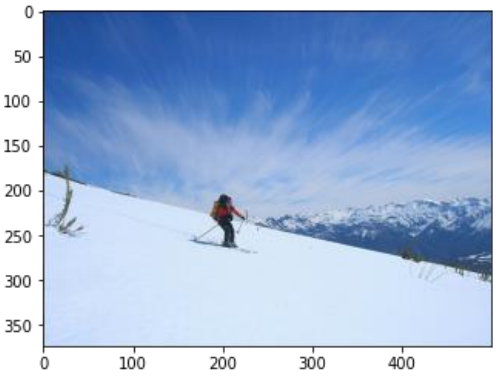

Fig (e) A trained dataset with caption “Man is standing on the edge of mountain”

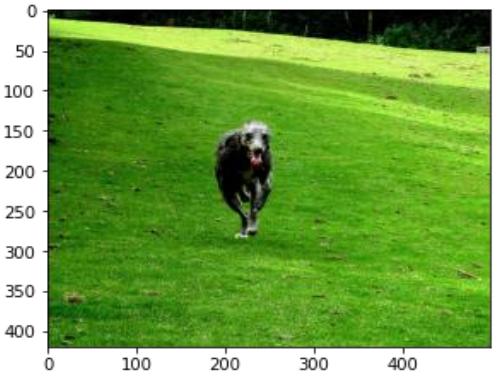

Fig (f) A trained dataset with caption “Dog runs through the grass”

Fig no-7 Output images with captions

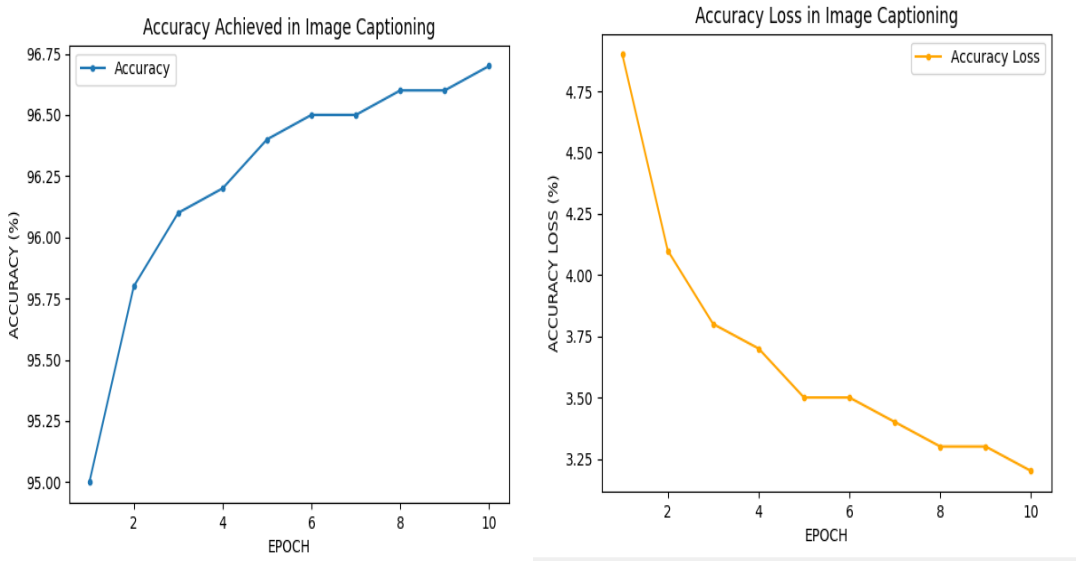

Fig (a) Maximum accuracy achieved by the model 96.75%" Fig (b) Maximum loss by the proposed model at 10 epochs

Fig no-8 Performance of the proposed model in terms of accuracy and loss function

Table no-1 Performance comparison of proposed model vs existing model

Noteworthy Contributions By Various Authors with Accuracy Rate

S.NO	AUTHOR	PUBLISHED YEAR	ACCURACY RATE
1.	Mareike Hartmann	2022	**95.46%**
2.	Mohamed Omri	2022	**93.47%**
3.	Georgious Zacharis	2022	**91.67%**
4.	Zeng	2022	**70-80.82%**
5.	Kumaravel Thangavel	2022	**75-82.43%**
6.	Suresh Muthusammy	2022	**70.48%**
7.	Jugal K. Kalita	2022	**75.66%**
8.	Rehab Ahmed	2022	**72.10%**
9.	Yulin Zhu	2022	**80.30%**
10.	Mayeesha Humaira	2021	**81.97%**
11.	Mingkui Tan	2021	**90.79%**
12.	Ariyo Oluwasammi	2021	**81.34%**
13.	Dora Zhao	2021	**90.79%**
14.	Yu Maemura	2021	**69.67%**
15.	Abdelrahman Mohammed	2020	**74.67%**
16	Proposed model	2022	**96.75%**

5. CONCLUSION AND FUTURE SCOPE

The evolution of image captioning over the last several years has been phenomenal. Recent years of study using deep learning have led to a rise in the precision of picture captioning. The textual description of the picture may improve the efficiency of content-based image retrieval, opening up new possibilities for the use of visual understanding in fields such as medicine, security, and the military. Image annotation, visual question answering (VQA), cross-media retrieval, video captioning, and video dialogue all have substantial academic and practical relevance and may benefit from the theoretical foundations and research approaches of image captioned. Following conclusion can be drawn from the proposed research work:

- The CNN-RNN as an encoder and decoder instead of CNN-CNN, RNN-RNN gives the best result as the present result witnessed so.
- The Algorithms such as VGG 16 and LSTM performs extremely well as it could help the model perform better and to achieve highest accuracy.
- The proposed model achieved maximum accuracy which is 963.75% and much higher than the many of the existing model.
- The no of epoch considered for the proposed model building was 10 which seem to be good while the loss function was 0.22.

REFERENCES

1. Mareike Hartmann et al., (2022),” Interactive Machine Learning for Image Captioning”, arXiv.
2. Mohamed Omri et al.,(2022),” Modeling of Hyperparameter Tuned Deep Learning Model for Automated Image Captioning” , Mathematics 2022, 10, 288.
3. Deepak Mathur, N. K. V. . (2022). Analysis & Prediction of Road Accident Data for NH-19/44. International Journal on Recent Technologies in Mechanical and Electrical Engineering, 9(2), 13–33. https://doi.org/10.17762/ijrmee.v9i2.366
4. Georgios Zacharis et al .,(2022),” Implementation and Optimization of Image Processing Algorithm using Machine Learning and Image Compression”, SHS Web of Conferences 139.
5. Kumaravel Thangavel et al., (2022),”A novel method for image captioning using multimodal feature fusion employing mask recurrent neural networks and long short term memory”, Research Square.
6. Chauhan, T., and S. Sonawane. “The Contemplation of Explainable Artificial Intelligence Techniques: Model Interpretation Using Explainable AI”. International Journal on Recent and Innovation Trends in Computing and Communication, vol. 10, no. 4, Apr. 2022, pp. 65-71, doi:10.17762/ijritcc.v10i4.5538.
7. Abdelrahman Mohamed et al., (2022),” Image Captioning Through Self-Supervised Learning”, Technical Report ·
8. Rehab Alahmadi et al., (2022,” Improve Image Captioning by Estimating the Gazing Patterns from the Caption”, WACV, Computer Vision Foundation.
9. Pang-Jo Chun et al., (2022),” A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage”, Comput Aided Civ Inf. 2022.

10. Yulin zhu et al.,(2022),” Image-Based Storytelling Using Deep Learning”, The 5th International Conference on Control and Computer Vision (ICCCV 2022).
11. Gill, D. R. . (2022). A Study of Framework of Behavioural Driven Development: Methodologies, Advantages, and Challenges. International Journal on Future Revolution in Computer Science &Amp; Communication Engineering, 8(2), 09–12. https://doi.org/10.17762/ijfrcsce.v8i2.2068
12. Zanyar Zohourianshahzadi & Jugal K. Kalita (2021),” Neural Attention for Image Captioning: Review of Outstanding Methods”, Springer Nature Artificial Intelligence.
13. Mayeesha Humaira et al., (2021,” A Hybridized Deep Learning Method for Bengali Image Captioning”, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 12.
14. J. . Hermina, N. S. . Karpagam, P. . Deepika, D. S. . Jeslet, and D. Komarasamy, “A Novel Approach to Detect Social Distancing Among People in College Campus”, Int J Intell Syst Appl Eng, vol. 10, no. 2, pp. 153–158, May 2022.
15. Ariyo Oluwasammi et al., (2021),” Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning”, Hindawi Complexity Volume 2021,
16. Gill, R. . (2022). Subalgebras over p-Adic Domains. International Journal on Recent Trends in Life Science and Mathematics, 9(1), 37–46. https://doi.org/10.17762/ijlsm.v9i1.140
17. Guanghui Xu et al.,(2021),” Towards Accurate Text-based Image Captioning with Content Diversity Exploration”,CVPR-2021.
18. Baohua Sun, et a., (2020),” SuperOCR: A Conversion from Optical Character Recognition to Image Captioning”, arxiv-2020.