IJNRD.ORG ISSN : 2456-4184
International Research Journal
IJNRD
Research Through Innovation
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

A COMPREHENSIVE SURVEY ON IMAGE CAPTIONING

[1]Manjiri Ranjanikar, [2]Shruti Biradar, [3]Rohan Dupade, [4]Vipul Gaikwad, [5]Harsh Pawar

[1,2,3,4,5]Department of Computer Engineering,
Pimpri Chinchwad College of Engineering, Pune, India

Abstract : An image captioning system uses both computer vision and natural language processing modules. While the computer vision module identifies important items or extracts information from images, the Natural Language Processing (NLP) module accurately constructs syntactic and semantic picture captions. Due to its significance in real-world applications like the analysis of enormous amounts of raw photos and the detection of previously unidentified patterns for machine learning applications used to drive self-driving cars and create software that helps the blind, a lot of people have recently developed an interest in the automatic generation of a natural language description or text of an image. It integrates computer vision with natural language processing, two essential components of artificial intelligence. In this study, image captioning will be done using neural networks. ResNet is used as the encoder to retrieve the image data, To create subtitles for videos that use the built-in language, CNN is used as an encoder to retrieve data from the picture and neural convolutional network RNN (Long-Short-Term Memory) as a decoder.

***IndexTerms* - Deep learning, Natural Language Processing, Image Captioning, Visual-language, ResNet, LSTM, Recurrent Neural Network, Convolutional Neural Network.**

I. INTRODUCTION

Image captioning, also known as photo captioning, is the process of automatically creating written descriptions for an image. The content of the image is described in the text. While computer vision is used to forecast attribute labels and other information about events and characters, natural language processing is utilized to create captions. This process encompasses the concepts of object identification, object relationship discovery, and attribute prediction. Every object has a few unique qualities that significantly describe it. The future of deep neural networks in the area of object detection is very promising. The ability to learn from previously completed samples and use that information to predict the outcome was demonstrated by deep neural network learning. Deep learning networks' output layer can predict labels and classifications since it is more abstract than the prior layers.

To keep the knowledge for a longer period, a captioning model combines CNN (Convolutional Neural Networks) and RNN (Recurrent Neural Networks) with LSTM (Long Short Term Memory), a specific kind of RNN with a memory cell.
The challenge must be described to develop an image caption generator utilizing the ResNet-LSTM model. The system employs convolutional neural networks for the visual identification of the image captioning model and recurrent neural networks for the language-based model. Convolutions are carried out using matrices built from an image's pixel-level data. Class names are provided based on the results of the output matrix, which locates a particular pattern in the source image. The features of the image were typically extracted using CNN and then linearly transformed into feature vectors. Then, to generate descriptions, the LSTM networks are used as a language model. CNN and LSTM will be used to construct an image caption generator.

II. LITERATURE REVIEW

Moses et. al.[1] creates a generative CNN-LSTM model that outperforms human baselines by 2.7 BLEU-4 points and is extremely comparable to state-of-the-art (3.8 CIDEr points lower). Studies using the MS COCO dataset set show that it typically results in accurate and understandable captions, and hyperparameter optimization using dropout and the amount of LSTM layers helps us to lessen the effects of overfitting. We also show that, despite having different prior contexts, words that are semantically close to one another when they are emitted—like "plate" and "bowl"—move the LSTM hidden state similarly, and that divergence in the hidden state only occurs when semantically distant words—like "vase" and "food"—are emitted. This gives the interplay between learned word embeddings and LSTM hidden states a semantic significance.

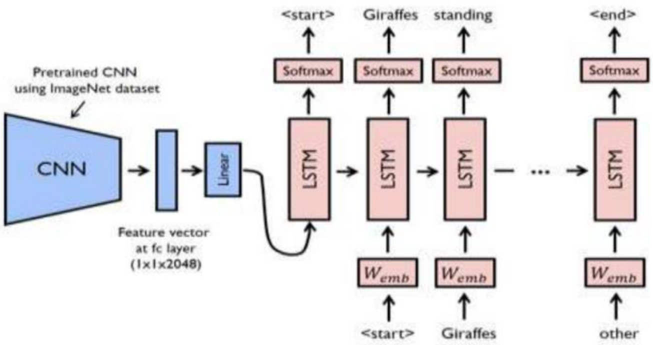


Fig 1: CNN-LSTM flow Diagram[1]

R. Subash et. al.[2] demonstrated directly contrasting the query picture with database photos, previous methods recovered relevant hopeful normal language states. However, although closed visual vocabularies can be useful for demonstration, they are incredibly limited in comparison to the vast amount of detailed images that a person is capable of creating. The variety of created inscriptions was pushed to its breaking limit by these approaches. The model should not make any assumptions about explicit predetermined formats, standards, or classifications; instead, it should rely on understanding how to construct sentences using the preparatory data. Convolution neural networks are used in the recommended model to help separate important components of the image from which the subtitle is to be built. After that, logical statements are formed and inscriptions are created utilizing a probabilistic technique and Natural Language Processing Techniques.

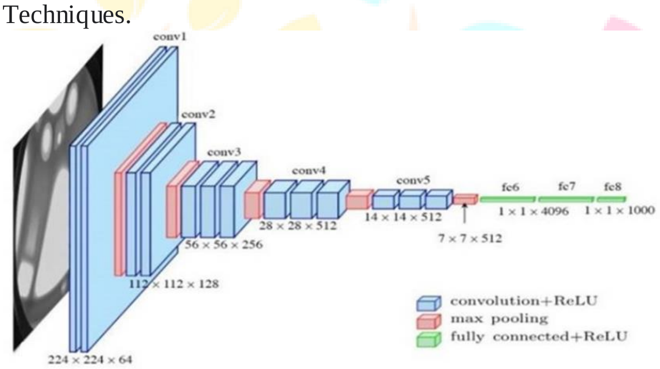
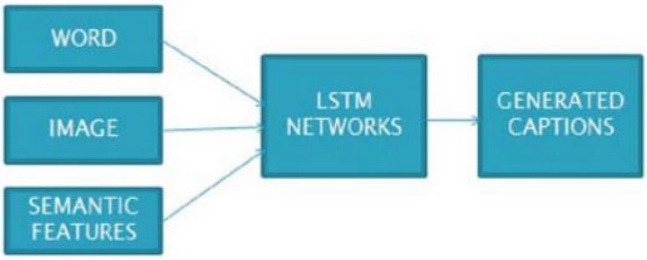


Fig 2: VGG16 Architecture[2]

In addition to the feature extraction technique, semantic keyword retrieval, face detection, and encoder-decoder LSTM networks, Abisha Anto Ignatious. L et. al.[3] recommended that the semantic-driven CNN-LSTM model be included in this study. To identify characteristics in an image, a pre-trained CNN is employed. The components in the image are recognized using an extraction module for semantic keywords. Using the image's semantic tags, the objects that have been identified are given names. The use of such semantic labels enhances its captions' capacity to appropriately describe the items. Each word of the captions is made by LSTM-based word embedding. The faces dataset, which contains the facial photos of 232 celebrities, has been collected and used by the facial recognition system to identify and recognize famous faces in images. To create individual subtitles, the person's name was replaced with the text's instances of them. To determine the accuracy of the generated captions, Bilingual Evaluation Understudy (BLEU) and METEOR scores were established



.

Fig 3: Language Model Architecture[3]

Himanshu Sharma et. al.[4] suggested a technique for initiating the images that makes use of both visual and extra information from information sources like ConceptNet.Using Flickr8k and Flickr30k, two publicly accessible datasets, they showed the method's utility. The results demonstrate that the proposed model performs more effectively than the state-of-the-art techniques for generating image captions. Finally, we will discuss potential future developments in image captioning.

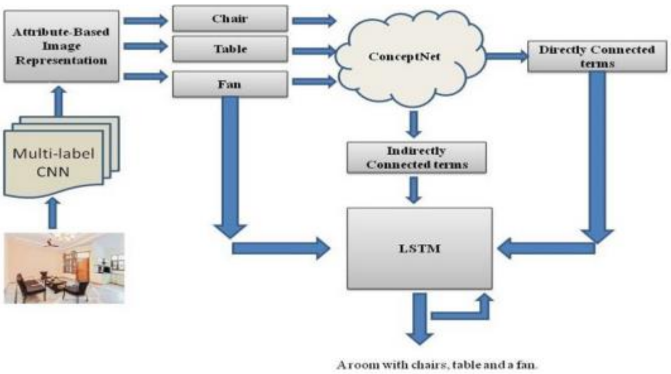

Attribute-Based Image Representation
Chair
Table
Fan
ConceptNet
Directly Connected terms
Multi-label CNN
Indirectly Connected terms
LSTM
A room with chairs, table and a fan.

Fig 4: The proposed image captioning incorporating external knowledge[4]

G. Geetha et. al.[5] proposed an algorithm to aid in the global community's comprehension of where, how, and why deforestation occurs on a worldwide scale. Future advancements in satellite imaging technology will open up new possibilities for a more accurate analysis of both major and minor alterations to the planet, such as deforestation. Nearly a fifth of the Amazon rainforest has been destroyed in the past 40 years. This application was created to estimate and analyze the forestTo predict multiclass multi-label images, several classification frameworks are used, such as gate recurrent unit label captioning and sparse cross-entropy. Deep convolutional neural networks (CNNs) are trained using satellite imagery to learn visual attributes. By adjusting an architecture that combines a pre-trained VGG-19 encoder with a GRU decoder that was trained on ImageNet data.

Wang et al.[6] proposed the architecture of Bidirectional LSTM in 2016. Wang researched deep bidirectional LSTM architectures to enhance the output of the image captioning model and take advantage of both the past and the future information provided in the series during word translation. A CNN was utilized to extract the visual elements, and the deep Bi-LSTM architecture was employed to translate the text. According to their findings, the Bi-LSTM approach increased the formation of extended word sequences since it could produce successive words more effectively, which was reflected in greater BLEU.

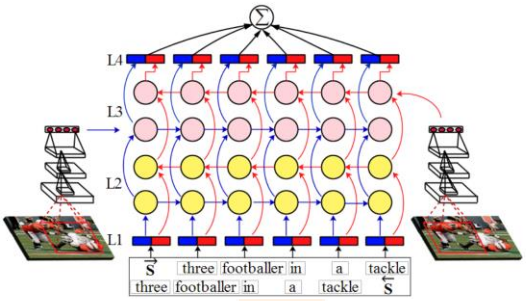


Fig 5: Multimodal Bidirectional LSTM[6]

Fig: 5 shows a different layer where L1 is called the sentence embedding layer. The T-LSTM layer is L2. M-LSTM layer is the name given to L3i. And the last is the Softmax layer. By feeding sentences both in the forward direction, as shown by the blue arrow, and the backward direction, as shown by the red arrow, this model may produce sentences word by word over time. By minimizing a joint loss, our model is trainable from beginning to end.

Bin Yi et. al.[7] proposed a novel framework for video captioning that combines soft attention and bidirectional long-short-term memory (BiLSTM) to improve global representations of videos and the capacity to recogni ze persistent emotions in them.

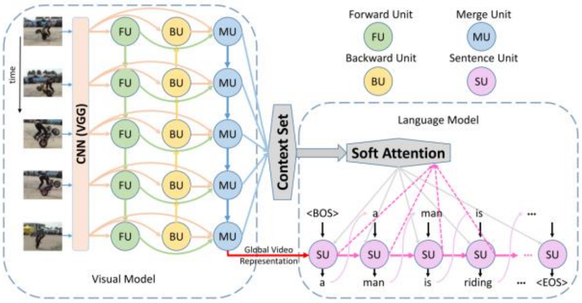


Fig 6: Model Flowchart for video captioning[7]

Thus, the basic implementation of the visual encoder and natural-language generator components was done using two LSTM networks. Then, they focused on the initial CNN characteristics, which were capable of precisely capturing the bidirectional global temporal structure in video clips, and employed a bidirectional LSTM network to encode video sequences. The sentences in the video were then converted to text using a language decoder. Experimental results show that they outperformed numerous other state-of-the-art approaches on the MSVD and MSR-VTT 10K datasets.

Md Zakir et. al.[8] for image captioning, they recommended the Bi-Directional Self-Attention method. In several sequence modeling problems, a novel attention mechanism known as "self-attention" has gained popularity. This technique is attention-based and does not rely on LSTM/CNN. They employed bi-directional self-attention, which shifts focus backward and forward, to caption the pictures. They then computed attention in both directions using two masks, a front mask, and a backward mask. LSTMs cannot perform parallel computation because of temporal reliance, therefore to solve this problem, self-attention is used which is highly parallelized and does not require temporal dependency. To get a rich feature representation, bi-SAN and inter-attention can be effective methods. It doesn't need any repetition or convolution. As a result, it computes faster but also can record a sequence's distant dependencies. They have shown that the suggested Bi-SAN-based image captioning methodology outperforms traditional means on the ROUGH-L metrics and BLEU-1, 3, 4.

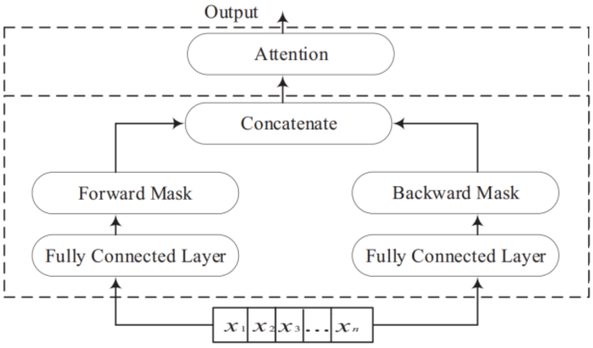


Fig 7: Sequence modeling with self-attention in both directions[8]

Fig 7 shows the bi-directional self-attention model which is been used in the sequence modeling in the proposed model [8]

Xinxin et. al.[9] designed a revolutionary Check Image Captioning Job that needs reasoning about both past and future word order to rebuild relevant image descriptions. They demonstrate the efficacy of their strategy using both this task and the Visual Madlibs dataset, consistently surpassing all baseline techniques. They provided the first approximation inference approach for bidirectional neural sequence models' 1-Best (and M-Best) decoding. To "fill in the gaps," these algorithms use known knowledge from the past and the future. They create a bidirectional beam search method that is fast to calculate in bidirectional recurrent neural network models based on an estimation of the full joint distribution over output sequences of the image.

Stefan Lee et. al. [10] developed a model for captioning images using stack parallel LSTM and triple attention. There were three different stages of LSTM, The attention model is used as the LSTM model's input in the first stage, the LSTM model and attention model are combined in the second stage, and the attention model is used as the LSTM model's output in the third stage. By evaluating the model on the MS-COCO dataset, we can show that the triple attention (TA- LSTM) method, which uses visually relevant information at every phase of the LSTM process, enhances the model's performance in comparison to traditional LSTM.

In the approach proposed by Ghadah Alabduljabbar et al.[11], the model is developed overall to perform well with minimal engineering work. Unlike other approaches that focus just on visual information and use structural or worldwide information, our approach makes use of the fine-grained visual data obtained using ResNet101. Utilizing a refining model, a clean representation of the retrieved object features is made[11].

.

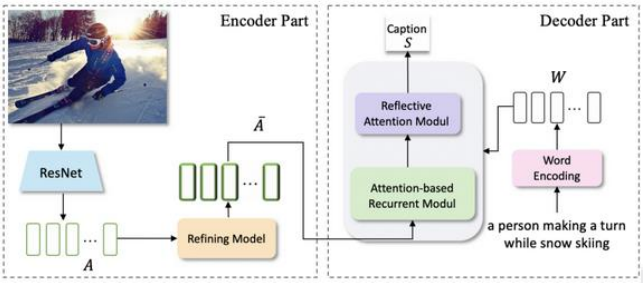

Encoder Part
Decoder Part
ResNet
A
Refining Model
$\bar{A}$
Caption
S
Reflective Attention Modul
Attention-based Recurrent Modul
W
Word Encoding
a person making a turn while snow skiing

Fig 8: A summary of the recommended architecture.

In the method put forth by Yuqing Peng et al.[12], the problem of the present image semantic comprehension model, which is incorrect or overlooks scene identification, and has a low accuracy rate for the description sentence, has been addressed by the proposal of an image semantic comprehension model combining scene components. The LDA analysis corpus's text volume is used by this model to determine the theme first.
The global characteristic of the picture is then extracted using ResNet, while the deep scene feature is extracted using Places365-CNNs. The model also makes use of the scene data from the corpus and the images.

Aesthetic Image Captioning (AIC) is a specialized end-to-end network. This method was proposed by Xin Jin et al.[13] uses photographs and associated aesthetic judgments for training and generates textual interpretations of visual viewpoints on aesthetics. A model that generates sequences frequently uses ResNet101 + LSTM, where an encoder permanently encodes the data and a decoder word-by-word decodes it into a sequence.

Utilizing the ImageNet classification challenge, the pre-trained ResNet-101 model is trained following the suggested technique by Xianrui Li et al.[14] the attribute detection model known as AttriNET is built on the Pre-trained ResNet-101. The transfer learning framework is built on top of the pre-trained ResNet-101. For the same, a framework for combined attribute detection and visual attention is used.

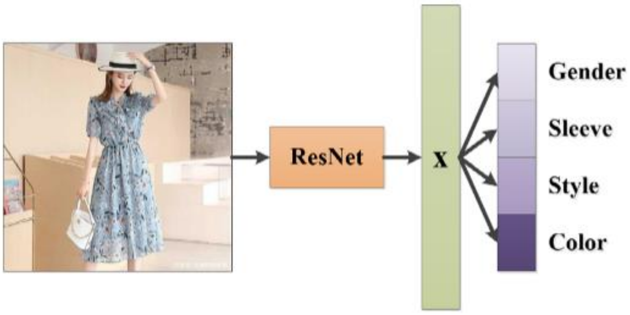


Fig 9: AttriNet for detecting clothing characteristics [14]

In the approach put out by Shobiya L, Pradheesha R, and Prof. Kala[15], natural sentences describing an image are produced using a generative model that combines recent development
in the vision of computers and the translation of machines. The components used in the method are analyzed as Residual Networks (ResNet), RNN, and the generation of sentences
using LSTM. The model's performance is evaluated using the standard evaluation of the matrix which outperforms the previous benchmark models.

Aishwarya Maroju et. al.[16] proposed a technique in which LSTM is used for decoding and Resnet Architecture is utilized for encoding. The model is been trained using data from the Flickr8k dataset. When we use the Graphic Processing Unit to execute the model, it operates effectively. The Resnet-LSTM model is more accurate than the CNN-RNN model and the VGG model.

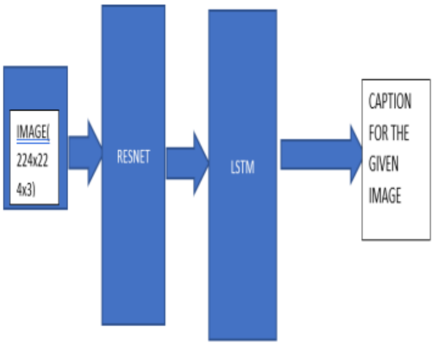
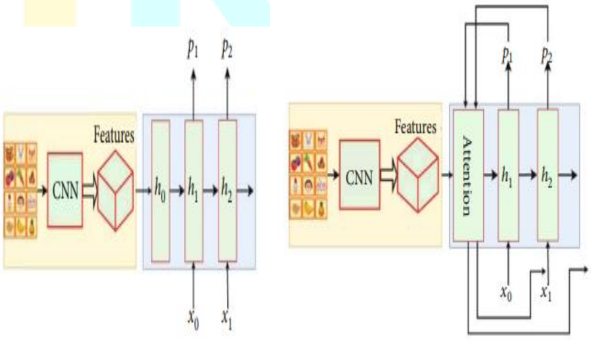


Fig 10:Resnet-LSTM Architecture[16]

Yan Chu et. al.[17] presented a joint model of AICRL. AICRL can carry out soft attention-based automatic picture captioning using ResNet50. An encoder plus a decoder make up AICRL. The model is more efficient overall when ResNet50 is used. Performance on all metrics, including METEOR, CIDER, and BLEU-4, improves while using the soft attention technique.



(a) (b)

Fig 11: Model without(a) and with attention(b)[17]

Shruti Mundargi et. al.[18] used the Resnet-LSTM model for image captioning. LSTM is used for decoding and Resnet is used for encoding. This model performs well with the aid of the graphics processing unit, the model was executed. Resnet50 reduces consumption costs and training time.

Janvi Jambhale et. al.[19] provided an analysis of the visual attention mechanism. Captions using three preprocessing models Inception V3, VGG19, and Resnet50 were analyzed.

The photos' crucial components were identified. Efficient results were obtained using the Ensemble learning technique.

A study emphasizing the impact of attention-based learning on automatically generated Hindi captions was presented by Abhishek Sethi et al.[20] in the publication. The Hindi-language Flickr8K dataset is utilized to verify how well the suggested work performs in terms of BLEU score. The methodology for the investigation includes a parallel architecture with two methods. The first method takes in the image data and generates characteristics using it. Numerous pre-trained, well-known CNN models were used, such as Resnet50 and VGG16. Processed reference captions are fed into an LSTM by the second pipeline. The output data from both streams are combined to produce captions for the image using a combination of various thicknesses composed of 256 neurons.

III.DISCUSSION

3.1. Resources for Computation and Model Architecture

The popular encoder-decoder framework, which is adaptable and powerful, is used by the majority of the models. It is referred to occasionally as a CNN+RNN structure. Recurrent neural networks (RNNs) often serve as the decoder, whereas convolutional neural networks (CNNs) typically serve as the encoder. The encoder serves as the "reader" of the input image by extracting high-level deep features. The decoder is the component that generates words; given the encoded input images provided by the encoder, it generates words to convey the image in a comprehensive, grammatically correct phrase.

3.1.1 Image encoding by CNN

CNNs are used for image encoding in all deep-learning image captioning models due to their propensity to spot patterns in the intensity values of individual pixels in a picture. The Image Encoding module mostly implements the AlexNet, AlexNet, VGGNet, VGGNet, GoogLeNet/Inception-V1, ResNet, Inception V3, and Inception V1 CNN architectures. VGGNet is the earliest and most widely used option for feature extraction from photos and is chosen for both its simplicity and strength. Though it was also discovered that ResNet was used as an encoder in the same number of publications. ResNet is the most computationally efficient convolutional network available.

Tabel 1: Comparison chart for table architecture

Architecture	Year	#Multiply-Adds	#Params	Top-1 Accuracy	Top-5 Accuracy
Resnet-50	2015	3.9B	25.5 M	75.2	93
Inception-V1	2013	1.43B	7M	69.8	89.3
VGG	2013	15.5B	138M	70.5	91.2
Alexnet	2012	724M	61M	57.1	80.2

3.1.2 LSTM for decoding

The term "LSTMs," or Long Short-Term Memory Networks, refers to a certain kinds of RNNs that can identify long-term correlations. The hardest problems, such as speech recognition, machine translation, and many others, are where technology is most

frequently applied. When training conventional RNNs, this issue was discovered because, as a neural network is advanced, if the gradient is very small or zero, very little to no training can occur, resulting in poor forecasting performance. Given that there may

be delays of various lengths between important occurrences in a time series, LSTM networks are well suited for categorizing, analyzing, and predicting outcomes based on the dataLSTM outperforms RNN's short-term memory restrictions in terms of

Performance, making it substantially more effective and superior. The LSTM can process inputs while processing pertinent information, and it can ignore irrelevant information.

3.2 Datasets

1. Flickr30k, which comprises 31,783 total images, is a popular dataset for image captioning. It takes into consideration the distribution of 29,783 images for training the model, 1,000 images for testing, and 1,000 images for validation. There are five captions for each image in this collection.

2. MS COCO comprises one of the datasets that are most frequently used for image captioning. There are 123,287 photos in all. The dataset for this study was split into 113,287 photos for training the corresponding models, 5,000 images for validation, and 5,000 images for testing. Every image has five captions that were created by humans just for that image.

Table 2. An analysis of the datasets

Dataset	Testing split	Validation split	Training split	Total images
MS COCO	41k	83k	83k	144k
Flickr30k	1k	1k	28k	30k

IV. COMPARATIVE STUDY

4.1 Results and conclusion of the existing system

Soh, Moses et. al. [1]- The projection of a sentence's hidden states into two dimensions is contained in the results. Despite being conditional on different picture vectors, the first sentence pair shows that emitted words with similar semantic meanings shift the hidden state in the same direction. The second pair of sentences demonstrates how the emitted sequence shifts the hidden dimension similarly, with the hidden state sentence representation only diverging after the sentence starts to explain variations in the images.

R. Subash et. al. [2]- A calculation known as a bilingual evaluation understudy (BLEU) evaluates the type of content that a computer has been able to comprehend. It was one of the key metrics to get a strong correlation with human judgment. On the MSCOCO data set, they compared our model to other models that are already in use and found that it performs well.

Abisha Anto Ignatious. L et. al. [3]- The suggested model is put into practice utilizing Keras and a Tensorflow backend. The featurey As a language model, the LSTM with 128 cells is employed. 60% of the data set is used for training, 20% for validation, and 20% for testing.

Himanshu Sharma et. al. [4]- The results show that the suggested model outperforms the most recent methods for creating image descriptions. They demonstrated the model's efficacy using the widely used, freely accessible Flickr8k and Flickr30k datasets.

G. Geetha et. al. [5]- Findings vary depending on CNN: The Attention-based technique relies on a variety of different convolutional neural networks to get its results. Here is the outcome of a few tests. The foundation of attention-based approaches is the convolutional feature of convolutional neural networks; these features are used throughout attention-based methods by various CNN architectures.

Wang et. al. [6]- Because the Bi-LSTM architecture was more efficient at producing consecutive words, as demonstrated by higher BLEU scores, longer word sequences were produced. The results showed that the suggested bidirectional LSTM method optimized the generated captions. A bidirectional LSTM model proves that multi-task learning is beneficial to increase model performance.

Bin Yi et. al. [7]- Experiments on the MSR-VTT 10K and MSVD datasets demonstrated that they outperformed several other cutting-edge methods. Better global representations for videos can be created using a BiLSTM and a soft attention mechanism, which can also improve the detection of long-lasting motions in videos.

Hossain, Md Zakir et. al. [8]- Both forward and backward attention is been done. It achieves great performance on par with cutting-edge techniques. It Interattention and bi-SAN can be effective strategies for obtaining a rich feature representation without the need for recurrence or convolution. Consequently, it requires less computing time.

Zhu, Xinxin et. al. [9]- With results, they observe that the performance gets better than the base models. The test result is produced using a single model rather than the ensemble process. The triple attention mechanism and the LS-LSTM can produce sentences that are more legible and fluent, according to experiments.

Qing et. al. [10]- Developed a novel Bidirectional Beam Search (BiBS). Unidirectional RNNs generate illogical outputs because they are unable to construct meaning of both previous and upcoming outputs. For Bidirectional RNNs, this is the first top-B MAP inference approach.

Ghadah Alabduljabbar et. al. [11]- Instead of simply focusing on increasing visual features as in earlier methodologies, this model uses the extracted characteristic features and an attention-on-attention process to describe the interactions among elements in the image.

Yuqing Peng et. al. [12]- The study of the experiment's data shows that this method's double Long short-term memory fusion model and scene factors focus more on the impact of scene on total semantic meaning. By contrasting the model's experiment findings with those from the references, it has been established that the model presented here is accurate.

Xin Jin et. al. [13]- A new dataset of image aesthetics has been constructed. An image aesthetics-based latent Dirichlet assessment model was used to filter the constructed dataset. Finally, FAE-Captions, a new picture aesthetics dataset, is acquired. A novel convolutional neural network model that can generate ideal aesthetic comments demonstrates the effectiveness of the suggested model and the advantages of the data set.

Xianrui Li et. al. [14]- The proposed captioning for the clothing image effectively captures the image's characteristics and demonstrates the value of the work. With the help of a joint attribute detection and visual attention framework, the generation of captions is done.

Shobiya L et. al. [15] - The proposed model was built using a single encoder-decoder architecture. A CNN named ResNet101 was used as the encoder to turn a picture into a mini representation with graphical elements. The decoder for the descriptive sentence was then chosen as a language model LSTM. In the meanwhile, we combined the LSTM with the sensitive attention model so that learning may be targeted at a specific area of the image to enhance performance.

Aishwara Maroju et. al. [16]- Results of the suggested solution have demonstrated that the accuracy was relatively poor in the early periods. The captions generated were related if the model was run for more than 20 epochs. The accuracy was improved for 50 eras, and the captions were more accurate.

Yan Chu et. al. [17]- ResNet50 and LSTM-based automatic picture captioning methodology is offered. One encoding and one decoding device make up the AICRL model. The CNN uses the ResNet50, which aids in the generation of a vector with a fixed length. The following sentence is predicted by the decoder using the LSTM, which concentrates on specific areas of the image. The MS-COCO dataset, which aids in the intended descriptive sentence, was utilized to prepare the AICRL model.

Shruti Mundargi et. al. [18]- Results from the suggested solution indicate that the crucial elements of the photos were recognized. They employed the essemble learning technique, which combines the output of two or more machine learning models.

Janvi Jambhale et. al. [19]- A Res-Net-LSTM neural network-based model that can recognize objects in photos and produce descriptions for them is introduced. In this case, LSTMs are employed for decoding whereas Res-Net structure is used for encoding. When a picture is uploaded to Res-Net, it first extracts the image's features before training the model with these two inputs. Vocabulary is constructed using training caption data. After the model definition and fitting, the model was trained over 50 iterations. Early training epochs were found to provide captions that were not substantially related to the provided test images and had rather low accuracy. If the model has been trained for at least 20 epochs, the generated captions have some relationship to the given test images. They observe an increase in model accuracy and captions that are closely related to the test images after 50 epochs of model training.

Abhishek Sethi et. al. [20]- The results from the suggested model were not accurate as perversions of the English language. They used the flickr8k dataset, a well-known dataset for image captioning, as well as a Hindi version of the dataset that contains English captions in Hindi. The conversion cannot be completely precise because the grammar and vocabulary of the English and Hindi languages differ significantly. The model is trained using the provided captions, and it produces results appropriately.

V.CONCLUSION

Automated algorithms for detecting precancerous and cancerous lesions utilizing pictures collected during VIA are particularly promising for countries where expensive instruments and qualified workers are few, and where the majority of cervical cancer fatalities occur. A CNN-LSTM architecture has numerous uses in computer vision and natural language processing. Extremely powerful CNN networks can be used simultaneously for sequential data, like natural language. Resnet-LSTM is used to predict captions for the images. The Flickr 8k dataset's photos have anticipated captions. The model can automatically analyze an image and offer suitable captions in languages like English. Comparing ResNet-LSTM to CNN-RNN and VGG Models, the former is more accurate. This model performs well when it is run on the Graphic Processing Unit. The future scope for image captioning is the model for picture captioning powered by AI: It is an automated application that efficiently creates brief and insightful descriptions for massive amounts of photos. The model extracts detailed textual information about the provided photos using Natural Language Processing (NLP) techniques and computer vision.

REFERENCES

[1] Soh, Moses. "Learning CNN-LSTM architectures for image caption generation." Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep 1 (2016)

[2] Subash, R., R. Jebakumar, Yash Kamdar, and Nishit Bhatt. "Automatic image captioning using convolution neural networks and LSTM." In *Journal of Physics: Conference Series*, vol. 1362, no. 1, p. 012096. IOP Publishing, 2019.

[3] Ignatious, L. Abisha Anto, S. Jeevitha, M. Madhurambigai, and M. Hemalatha. "A Semantic Driven CNN–LSTM Architecture for Personalised Image Caption Generation." In *2019 11th International Conference on Advanced Computing (ICoAC)*, pp. 356-362. IEEE, 2019.

[4] Sharma, Himanshu, and Anand Singh Jalal. "Incorporating external knowledge for image captioning using CNN and LSTM." *Modern Physics Letters B* 34, no. 28 (2020): 2050315.

[5] Geetha, G., T. Kirthigadevi, G. Godwin Ponsam, T. Karthik, and M. Safa. "Image captioning using deep convolutional neural networks (CNN)." In *Journal of Physics: Conference Series*, vol. 1712, no. 1, p. 012015. IOP Publishing, 2020.

[6] Wang, Cheng, Haojin Yang, Christian Bartz, and Christoph Meinel. "Image captioning with deep bidirectional LSTMs." In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 988-997. 2016.

[7] Bin, Yi, Yang Yang, Fumin Shen, Ning Xie, Heng Tao Shen, and Xuelong Li. "Describing video with attention-based bidirectional LSTM." *IEEE transactions on cybernetics* 49, no. 7 (2018): 2631-2641.

[8] Hossain, Md Zakir, Ferdous Sohel, Mohd Fairuz Shiratuddin, Hamid Laga, and Mohammed Bennamoun. "Bi-SAN-CAP: Bi-directional self-attention for image captioning." In *2019 Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1-7. IEEE, 2019.

[9] Zhu, Xinxin, Lixiang Li, Jing Liu, Ziyi Li, Haipeng Peng, and Xinxin Niu. "Image captioning with triple-attention and stack parallel LSTM." *Neurocomputing* 319 (2018): 55-65.

[10] Sun, Qing, Stefan Lee, and Dhruv Batra. "Bidirectional beam search: Forward-backward inference in neural sequence models for fill-in-the-blank image captioning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6961-6969. 2017

[11] Alabduljabbar, Ghadah, Hafida Benhidour, and Said Kerrache. "Image Captioning based on Feature Refinement and Reflective Decoding." *arXiv preprint arXiv*:2206.07986(2022).

[12] Peng, Yuqing, Xuan Liu, Weihua Wang, Xiaosong Zhao, and Ming Wei. "Image caption model of double LSTM with scene factors." Image and Vision Computing 86 (2019): 38-44.

[13] Jin, Xin, Jianwen Lv, Xinghui Zhou, Chaoen Xiao, Xiaodong Li, and Shu Zhao. "Aesthetic image captioning on the FAE-Captions dataset." Computers and Electrical Engineering 101 (2022): 107866.

[14] Li, Xianrui, Zhiling Ye, Zhao Zhang, and Mingbo Zhao. "Clothes image caption generation with attribute detection and visual attention model." Pattern Recognition Letters 141 (2021): 68-74.

[15] Shobiya, L., and R. Pradheesha. "ImageCaption Generator Using RESNET-LSTM" *International Journal of Research in Engineering and Science (IJRES)* ISSN (Online): 2320-9364.

[16] Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati, 2021, "Image Caption Generating Deep Learning Model", *International Journal of Engineering Research & Technology (IJERT)* Volume 10, Issue 09 (September 2021).

[17] Chu, Yan, Xiao Yue, Lei Yu, Mikhailov Sergei, and Zhengkui Wang. "Automatic image captioning based on ResNet50 and LSTM with soft attention." *Wireless Communications and Mobile Computing* 2020 (2020): 1-7.

[18] Mundargi, Shruti, and Hrushikesh Mohanty. "Image Captioning using Attention Mechanism with ResNet VGG and Inception Models." *International Research Journal of Engineering and* Technology *(IRJET)* 7, no. 09 (2020).

[19] Jambhale, Janvi, Shreeya Sangale, Aarti Avhad, Payal Vairagade, and Jameer Kotwal. "Image caption generator using convolutional neural network and long short-term memory." *International Research Journal of Modernization in Engineering Technology and Science*(2022).

[20] Sethi Sethi, Abhishek, Aditya Jain, and Chhavi Dhiman. "Image Caption Generator in Hindi Using Attention." In *Advanced Production and Industrial Engineering*, pp. 101-107. IOS Press, 2022.