IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)
An International Open Access, Peer-reviewed, Refereed Journal

Integration Of Deep Learning And Natural Language Processing In Image Captions Generation

[1]Shivani Sharma, [2]Vinay Kumar
[1]M.tech, Student [2]Assistant Professor
[1]Dept. of CSE, B.N. College of Engineering and Technology
[2]Dept. of CSE, B.N. College of Engineering and Technology, Lucknow, India

Abstract: Image captioning is an interesting and challenging task with applications in diverse domains such as image retrieval, organizing and locating images of users' interest, etc. It has huge potential for replacing manual caption generation for images and is especially suitable for large-scale image data. Recently, deep neural network-based methods have achieved great success in the field of computer vision, machine translation, and language generation. In this research, we offer an encoder-decoder based model that can produce image captions with proper grammar. This model uses LSTM as a decoder and VGG16 Hybrid Places 1365 as an encoder. The model is trained on labeled Flickr8k and MS-COCO Captions datasets to guarantee full ground truth accuracy. Further, All widely used standard metrics, including BLEU, METEOR, GLEU, and ROUGE L, are used to evaluate the model. According to experimental findings, the suggested model achieved BLEU-1 score of 0.7350, METEOR score of 0.4768, and GLEU score of 0.2798 on the MS-COCO Caption dataset and BLEU1 score of 0.6666, METEOR score of 0.5060, and GLEU score of 0.2469 on the Flickr8k dataset. Comparing the suggested method to state-of-the-art techniques, a notable improvement in performance was thus obtained. We also present the outcomes of caption generation from real sample photographs, which support the validity of the suggested method and help assess the model's effectiveness even further.

***Index Terms* -** Neural network · Caption · CNN (Convolutional Neural Network) · Feature extraction · RNN (Recurrent Neural Network) · LSTM (Long Short-Term Memory)

I. Introduction

In computer vision, automatic picture caption generation is an ongoing research area. Because this issue combines two of the primary domains of artificial intelligence (AI), computer vision and natural language processing, and has a wide range of practical applications, researchers are drawn to it. Understanding an image is a prerequisite for creating a meaningful sentence out of it, and object identification and image classification can help with this.

With the significant advancements in Artificial Intelligence (AI), photos are now being used as input for a variety of functions. The study has addressed one application of AI. They identified the face using deep learning techniques. The primary goal of automatic image caption generation is to produce coherent sentences that explain the content of the image and the relationship between the items that are identified in the image. These words can then be utilized as suggestions in a variety of applications. It can be applied to a number of natural languages processing tasks, including social media recommendation, image indexing, virtual assistants, and visually impaired people. The creation of picture captions can aid machines in comprehending the content of images. It involves more than just finding objects in a picture; it also entails figuring out how the objects that have been found relate to one another. Image captioning techniques are divided into three categories by researchers: deep neural network-based, retrieval-based, and template-based. With template-based techniques, attributes, objects, and actions are first identified from the image, and then a number of blank slots in predefined templates are filled in. Retrieving an image that resembles the input image is how retrieval-based approaches generate captions.

While syntactically accurate captions are produced by these approaches, semantic accuracy and visual specificity cannot be guaranteed. Deep neural network-based techniques encode images first, then use a language model to generate captions. The deep neural network-based approach has the potential to produce semantically more accurate captions for the provided photos in comparison to the previous two methods. Most of the image relationship models in use today are built using deep neural networks. The first to define a multi-modal log-bilinear model for image captioning with a fixed context window was Kiros et al. CNNs and RNNs are primarily utilized in encoder-decoder image captioning models. In image-based tasks prior to the application of deep learning models, researchers typically employed Grey-level co-occurrence matrix (GLCM)and other machine learning-based feature extraction techniques. But in recent studies, CNN is used as an encoder to convert the input image into 1-D array representation, and RNN as a decoder or language model to generate the caption. Identification of proper CNN and RNN models is a challenging issue.

After reviewing the literature on picture caption creation, some important findings were discovered. Rather than using scene-specific models, the majority of state-of-the-art techniques employed CNN models that had been pre-trained on the object-specific ImageNet dataset. Consequently, these models produce object-specific captions. Second, the majority of the research that has already been done only provides an explanation of their findings using one or two evaluation criteria, including accuracy and BLEU-1 score. The following is a synopsis of our research contributions:

- This study proposes VGG16 Hybrid Places 1365 and LSTM as encoder and decoder respectively for automatic image captioning. The proposed model outperforms all state-of-art approaches.
- This study reports experimental results using all popular metrics such as BLEU, ROUGE L, METEOR, and GLEU on Flickr8k and MSCOCO Captions datasets.
- The study reports results of the proposed model on random live images for validation.

II. Related literature review

In this paper, authors surveyed the deep learning-based models used for image captioning on Flickr8k and MS-COCO datasets.

In template-based approach (Fig. 1 (a)), the process of caption generation is performed using predefined templates with a number of blank spaces that are filled with objects, actions, and attributes recognized in the input image. In the paper [3], the authors proposed the template slots for generating captions that are filled with the predicted triplet (object, action, scene) of visual components. Again, in the paper, the authors derived the objects (people, automobiles, etc., or things like trees, roads, etc.), qualities, and prepositions using a Conditional Random Field (CRF) based technique. The PASCAL dataset is used to evaluate the model using BLEU and ROUGE scores. The best ROUGE score in this work was 0.25, and the top BLEU score was 0.18. The paper's authors provided a way for creating new captions by carefully combining meaningful phrases from preexisting ones to produce a new one. A corpus of one million labeled images was employed.

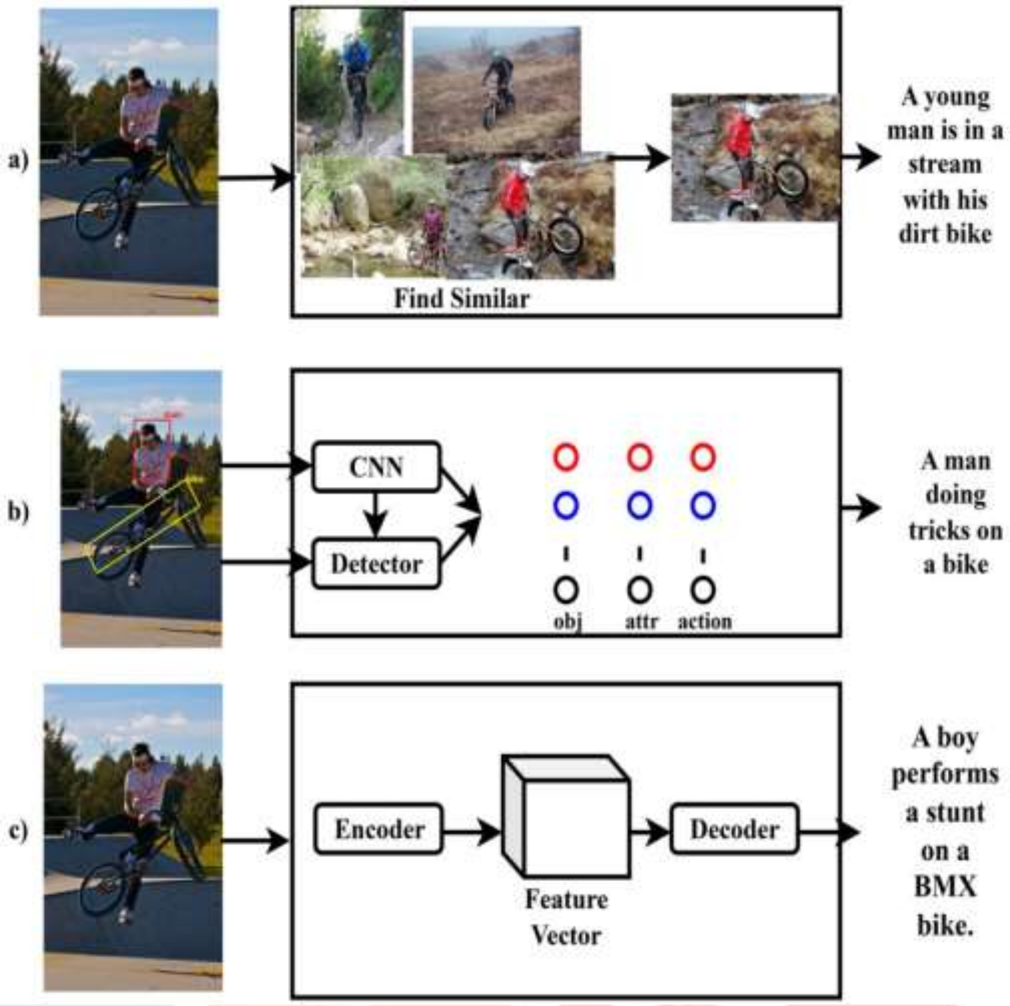

a)
Find Similar
A young man is in a stream with his dirt bike
b)
CNN
Detector
obj
attr
action
A man doing tricks on a bike
c)
Encoder
Feature Vector
Decoder
A boy performs a stunt on a BMX bike.

Fig. 1 Classification of image captioning models. (a) depicts template-based approach to image captioning, (b) shows retrieval-based captioning approach, (c) shows encoder-decoder based approach

dataset, with 1000 images put aside as a test set to compute BLEU (0.189) and METEOR (0.101) scores. These methods are basically hard to design and depend on pre-defined template.

The generation of captions for images in retrieval-based captioning systems (Fig. 1(b)) involves gathering visually comparable images. Once visually similar images are found, these kinds of techniques find captions for visually related photos from the training dataset and utilize those descriptions to deliver the caption of the query image. The authors of the research created a model for locating similar images among the many photographs in the dataset and returning the descriptions of these retrieved images to the query image based on millions of photos and their descriptions.

The text-based visual attention (TBVA) paradigm for automatically recognizing salient objects was proposed by the authors in the work [21]. They used the MS-COCO and Flickr30k datasets to evaluate the suggested model. The authors of the research [43] suggested a data-driven approach for retrieval-based image description creation. They came to the conclusion that the suggested strategy produced image captions in an effective and pertinent manner. These tactics yield generic, syntactically sound sentences, but they are unable to generate sentences that are image-specific and semantically right.

For picture captioning, the authors of the paper suggested dual graph convolutional networks with transformer and curriculum learning. They achieved a BLEU-1 score of 82.2 and a BLEU-2 score of 67.6 after evaluating the results on the MS-COCO dataset. The NIC (Neural Image Caption) model was put forth by the paper's authors and is based on encoder-decoder architecture. CNN is used as the encoder in this model, and its last layer is connected to the RNN decoder, which produces the text captions. LSTM is used as an RNN in this model.

Using a convolutional neural network (CNN) to encode an image into a numerical representation, the approach generates captions one word at a time by feeding the CNN's output into a decoder (RNN). Using the Flickr8K dataset, Yan Chu et al. proposed a model that combined ResNet50 and LSTM with soft attention, yielding a BLEU-1 score of 0.619. Two different model types were offered by Sulabh Katiyar et al.: an encoder-decoder model with attention and a basic encoder-decoder model. Using the Flickr8k dataset, these models produce BLEU-1 scores of 0.6373 and 0.6532, respectively.

In the work, a top-down and bottom-up R-CNN base method is suggested. Reranking the caption with beam search decoders and explanatory features improves the model. For automatic image caption generation, a Reference based Long Short Term Memory (R-LSTM) based approach is presented in the publication. To define pertinent captions, they employed a weighted technique that divided words and images. The Fliker30k and MS-COCO datasets were used to validate the suggested model, and the results of their experiment showed a 10.37% increase in the CIDr value on the MS-COCO dataset. To enhance the generated captions, a Hierarchical Refined Attention mechanism (CL-HRA) in a "Tell and Guess" Cooperative Learning model is developed.

The Cooperative Learning (CL) technique combines an image retrieval module (IRM) with an image caption module (ICM). The IRM is used to choose a picture from a collection of photos based on the description given, and the ICM is used to generate a descriptive natural language caption for a particular image. It was suggested to use a different CNN and LSTM based model to create captions for Geological Rock Images. The model focuses more on the image backgrounds in order to create captions from geological photographs of rocks.

III. METHODOLOGY

Artificial neural networks are used in deep learning, a subset of machine learning, to extract knowledge from data. A sizable dataset of photos and the captions that go with them is usually used to train the model when it comes to creating picture captions. The model gains the ability to link an image's visual characteristics to its associated textual description through training. The two primary parts of this design are the decoder (which generates sentences) and the encoder (which extracts features). The encoder takes an image and creates a feature vector that symbolizes the image's visual content. After receiving the feature vector, the decoder creates the caption's word sequence.

The methods used in the current investigation is covered in this section. The current study's primary goal is to generate well-written captions for input photographs. The following significant notions and ideas are pertinent in this regard.

3.1 Feature extraction

The encoder is used to extract the visual characteristics of an image. Generally speaking, convolutional neural networks (CNNs) are employed as encoders. They can extract semantic and content-based information from images and are frequently employed as models for visual identification tasks.

The first of them is the 16-layered VGG16 model [49], which on the ImageNet challenge yielded an accuracy of 92.7%. The attributes of the photos are provided by this model as a 1-D array, which is utilized for caption creation. When taught from scratch, the primary problem with VGG16 is its slow training speed. The second network, the Inception V3 [51] model, obtains an error rate of 4.2%, which is less than the VGG16 results that were previously published.

Comparing this network to the VGG Net, less processing power is needed. In comparison to VGG16 and Inception V3, the third and latest model, ResNet50, is utilized for deeper neural network learning. Table 1 displays the parameters and number of layers for each model.

3.2 Sentence generation

To produce sentences, the sentence generation component makes use of an RNN-based model. It is linked to the feature extraction model's output. Long word sequences are too difficult for the basic RNN to handle well. The Long Short-Term Memory (LSTM) network is used to address these problems. The most crucial part of the LSTM architecture is the memory cell.

Following its effective application in speech recognition, machine translation, and sequence learning, LSTM was also proven to be helpful in image captioning. LSTM is able to prevent vanishing and exploding gradient problems by using memory cells and gates. Three gates—the forget, output, and input gates—that are read and written by memory cell C are depicted in Figure 2. The LSTM takes inputs at time step I from a number of sources: Whereas Ci−1 denotes the previous memory cell state, Hi−1 represents the past hidden state, and Xi indicates the current input. The revised gate values at time step t for the inputs Xi, Hi−1, and Ci−1 are as follows:

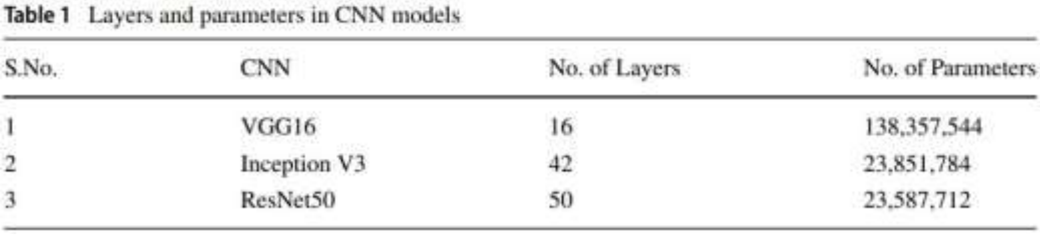


Table 1 Layers and parameters in CNN models

S.No.	CNN	No. of Layers	No. of Parameters
1	VGG16	16	138,357,544
2	Inception V3	42	23,851,784
3	ResNet50	50	23,587,712

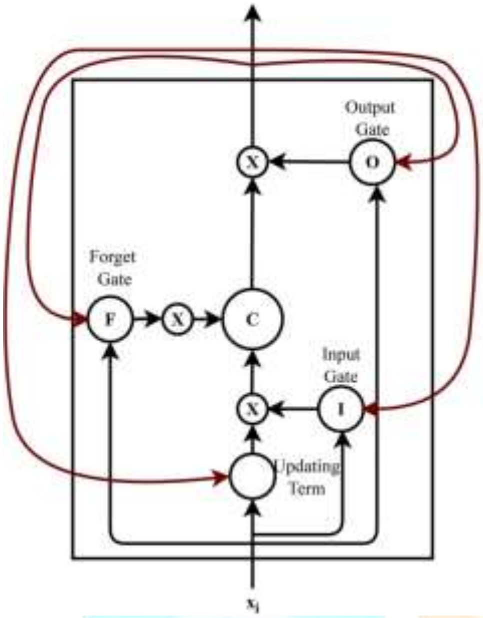


Fig. 2 Basic architecture of LSTM(Long Short Term Memory) cell

Ii = σ (WMXI Xi + WMH I hi−1 + BVI) (1)

Fi = σ (WMXF Xi + WMH F hi−1 + BVF) (2)

Oi = σ (WMXO Xi + WMHOhi−1 + BVO) (3)

Gi = φ(WMXCXi + WMHChi−1 + BVC) (4)

Ci = Fi ∗ Ci−1 + Ii ∗ Gi (5)

hi = Oi ∗ φ(Ci) (6)

Where XI , XF , XO, XC represent the inputs of the input gate, forget gate, output gate, and memory cell, WM represents weight metrics, and BV represents bias vectors. The sigmoid activation function is ρ and given as below.

$$\rho(X) = \frac{1}{1 + expo(-X)}$$

ϕ is the hyperbolic tangent, given as below.

$$\phi(X) = \frac{expo(X) - expo(-X)}{expo(X) + expo(-X)}$$

IJCRT

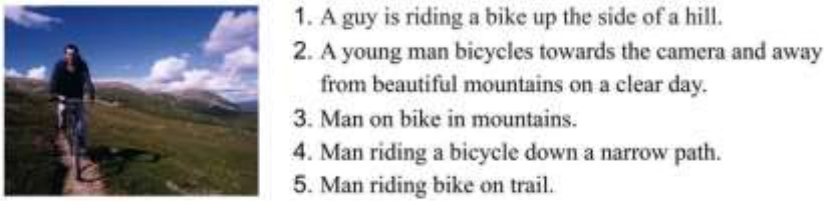
3.3 Dataset used

The well-known Flickr8k dataset was used in this study for training and performance assessment. Each image in this dataset has a personally tagged caption. The English-language captions for the photographs are included in the dataset. There are two components to the dataset: a description file and an image directory. There are 8000 photos in the image directory, and the description file has five subtitles for each picture.

Six thousand of the eight thousand photos were utilized for training, a further thousand for development, and the remaining thousand for testing. Figs. 3 and 6 display a few representative photos from the dataset with their English reference captions. Every picture is in jpeg format. The captions are twelve lines long on average. The input photos range in resolution from 256x500 to 500x500.

3.4 Proposed model

A deep neural network-based technique for creating image captions is shown in this subsection. As illustrated in Fig. 4, this method uses a recurrent neural network (LSTM) as a decoder and a convolutional neural network (VGG16 Hybrid Places1365) as an encoder-decoder to create captions for query photos.



1. A guy is riding a bike up the side of a hill.
2. A young man bicycles towards the camera and away from beautiful mountains on a clear day.
3. Man on bike in mountains.
4. Man riding a bicycle down a narrow path.
5. Man riding bike on trail.

Fig. 3 Sample image with reference captions

The VGG16 Hybrid Places1365 is used to generate the 1-D array representation of an image. Previously used CNN models such as VGG16, Inception, and ResNet are pretrained on the 1000-classes ImageNet dataset, whereas the VGG16 Hybrid Places1365 model is trained on both ImageNet and Places datasets (containing 1000 and 365 classes respectively).

Given image I , the captured visual features are denoted as a vector V = {v1, v2, ...vn}, which is calculated as:

V = CNNc(I)

where CNNc(I) is the output of the last convolutional layer of the modified model.

Following the effective training of the suggested model, the trained model predicts the caption for a given input image I word by word. The generated caption S has a probability of P (S/I) that is maximized.

With n words, the generated caption S becomes the sequence s1, s2,....., sn, where each word is a component of vocabulary V. During training, a search technique known as beam search is applied to provide captions for the images. The best n phrases are chosen for assessment in beam search as a set of length (t + 1) at a time t. In this work, a collection of five reference captions, or (n = 5), are used to train and test the model.

Using a K80 GPU and 12 GB of RAM, the Google Colab platform is used to train the suggested model. We used a number of libraries, including Tensorflow, Keras, pickle, OS, numpy, nltk, etc., to create the suggested model. Ten epochs of the model are run, with an average epoch duration of thirty minutes. For training purposes, 256 batch sizes and 3*3 kernel sizes are employed, respectively. The convolutional layer uses stride 1, and 0.5 and 0.003 are assumed to be the dropout value and leraning rate, respectively.

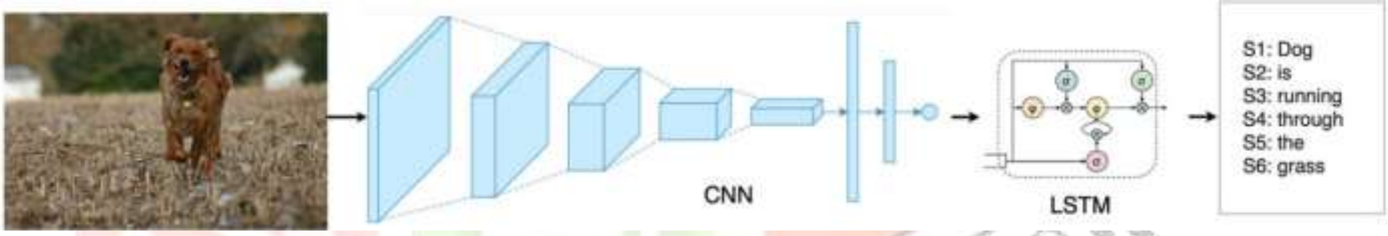
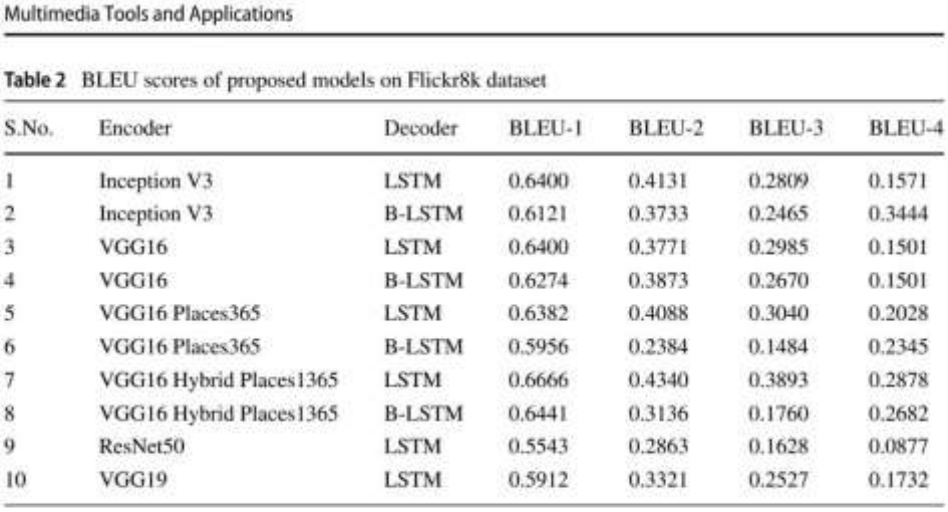


Fig. 4 Architecture of proposed model



Multimedia Tools and Applications

Table 2 BLEU scores of proposed models on Flickr8k dataset

S.No.	Encoder	Decoder	BLEU-1	BLEU-2	BLEU-3	BLEU-4
1	Inception V3	LSTM	0.6400	0.4131	0.2809	0.1571
2	Inception V3	B-LSTM	0.6121	0.3733	0.2465	0.3444
3	VGG16	LSTM	0.6400	0.3771	0.2985	0.1501
4	VGG16	B-LSTM	0.6274	0.3873	0.2670	0.1501
5	VGG16 Places365	LSTM	0.6382	0.4088	0.3040	0.2028
6	VGG16 Places365	B-LSTM	0.5956	0.2384	0.1484	0.2345
7	VGG16 Hybrid Places1365	LSTM	0.6666	0.4340	0.3893	0.2878
8	VGG16 Hybrid Places1365	B-LSTM	0.6441	0.3136	0.1760	0.2682
9	ResNet50	LSTM	0.5543	0.2863	0.1628	0.0877
10	VGG19	LSTM	0.5912	0.3321	0.2527	0.1732

Emphasis indicates result of best model

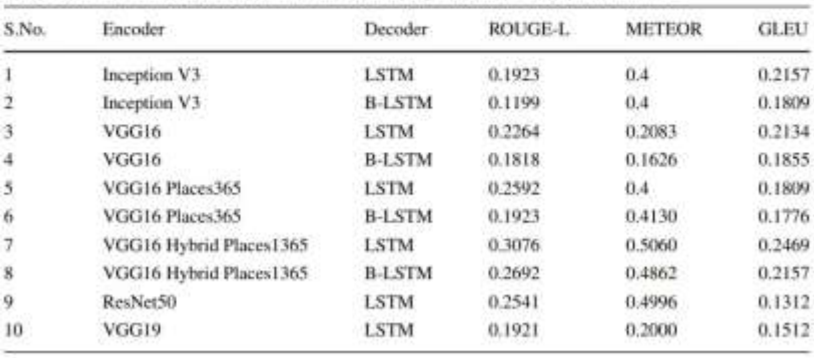
ReLU is the activation function utilized in the model's layers, and the "adam" optimizer is employed to reduce the "categorical crossentropy" loss.

4 Results and analysis

After the model has been trained, the resulting caption should have two desired qualities. First of all, it ought to make sense in relation to everything in the picture. Second, it ought to be intelligible and helpful to people. The suggested model produced a BLEU score of 0.6666, or 66.66%, on the Flickr8k dataset. The BLEU ratings derived from the Flickr8k dataset are displayed in Table 2 below.

After calculating the BLEU score, other relevant metrics including ROUGE L, GLEU, and METEOR were identified. Precision, recall, and F-Mean were acquired for the generated captions since they are required to calculate ROUGE L. The ROUGE L, GLEU, and METEOR scores on the Flickr8k dataset are displayed in Table 3 below.

Table 3 ROUGE.L, METEOR and GLEU scores of proposed models on Flickr8k dataset



S.No.	Encoder	Decoder	ROUGE-L	METEOR	GLEU
1	Inception V3	LSTM	0.1923	0.4	0.2157
2	Inception V3	B-LSTM	0.1199	0.4	0.1809
3	VGG16	LSTM	0.2264	0.2083	0.2134
4	VGG16	B-LSTM	0.1818	0.1626	0.1855
5	VGG16 Places365	LSTM	0.2592	0.4	0.1809
6	VGG16 Places365	B-LSTM	0.1923	0.4130	0.1776
7	VGG16 Hybrid Places1365	LSTM	0.3076	0.5060	0.2469
8	VGG16 Hybrid Places1365	B-LSTM	0.2692	0.4862	0.2157
9	ResNet50	LSTM	0.2541	0.4996	0.1312
10	VGG19	LSTM	0.1921	0.2000	0.1512

Emphasis indicates result of best model

10 distinct models were used in our experiments, as Tables 2 and 3 demonstrate. Using the Inception V3 with LSTM, which produced a BLEU-1 score of 0.64, the study began. Eventually, we created the VGG16 Hybrid Places 1365 model with LSTM after refining this baseline model. With a BLEU score of 0.6666, this model outperformed all the other models that were tested. Other widely used metrics, like ROUGE L, METEOR, and GLEU, which have respective values of 0.3076, 0.5060, and 0.2469, showed the similar pattern.
According to the experiment results, CNN using LSTM outperforms CNN using B-LSTM. It is likely that the B-LSTM does not produce better results for the selected dataset and experimental circumstances because of the increased number of parameters, which necessitates more adjustment to attain optimal performance.

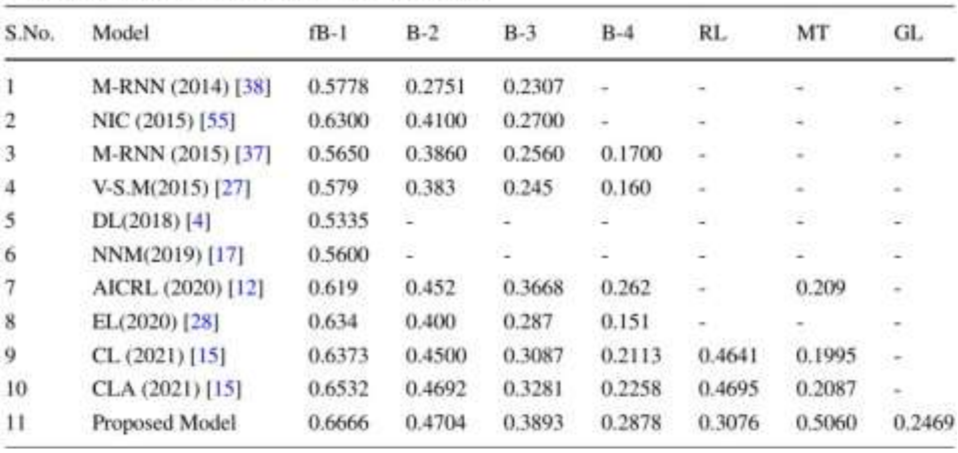
4.1 Comparison with state-of-the-art

We compared the outcomes with the most advanced models, as indicated in Tables 4 and 5, as well as Fig. 5, to demonstrate the effectiveness of the suggested model. The model's efficacy is corroborated by the outcomes on several metrics, including GLEU (0.2469 & 0.2798) and METEOR (0.5060 & 0.4768). It is significant to highlight that the majority of cutting-edge studies do not disclose their findings using the aforementioned measures.

4.2 Discussion and Limitations of proposed model

As elaborated in Section 4.2 and presented in Tables 4 & 5, the suggested model yields significantly superior outcomes when compared to cutting-edge techniques on both datasets (MS Coco & Flickr8k). Additionally, the suggested model's effectiveness is assessed using a random sample of live images, as demonstrated in Fig. 6.

Table 4 Performance comparison of proposed model against state-of-the-art techniques on Flickr8K dataset (- indicates that the result is not reported in the paper, B-1: BLEU-1, B-2: BLEU-2, B-3: BLEU-3, B-4: BLEU-4, RL: ROUGE.L, MT: METEOR, GL: GLEU)



S.No.	Model	fB-1	B-2	B-3	B-4	RL	MT	GL
1	M-RNN (2014) [38]	0.5778	0.2751	0.2307	-	-	-	-
2	NIC (2015) [55]	0.6300	0.4100	0.2700	-	-	-	-
3	M-RNN (2015) [37]	0.5650	0.3860	0.2560	0.1700	-	-	-
4	V-S.M(2015) [27]	0.579	0.383	0.245	0.160	-	-	-
5	DL(2018) [4]	0.5335	-	-	-	-	-	-
6	NNM(2019) [17]	0.5600	-	-	-	-	-	-
7	AICRL (2020) [12]	0.619	0.452	0.3668	0.262	-	0.209	-
8	EL(2020) [28]	0.634	0.400	0.287	0.151	-	-	-
9	CL (2021) [15]	0.6373	0.4500	0.3087	0.2113	0.4641	0.1995	-
10	CLA (2021) [15]	0.6532	0.4692	0.3281	0.2258	0.4695	0.2087	-
11	Proposed Model	0.6666	0.4704	0.3893	0.2878	0.3076	0.5060	0.2469

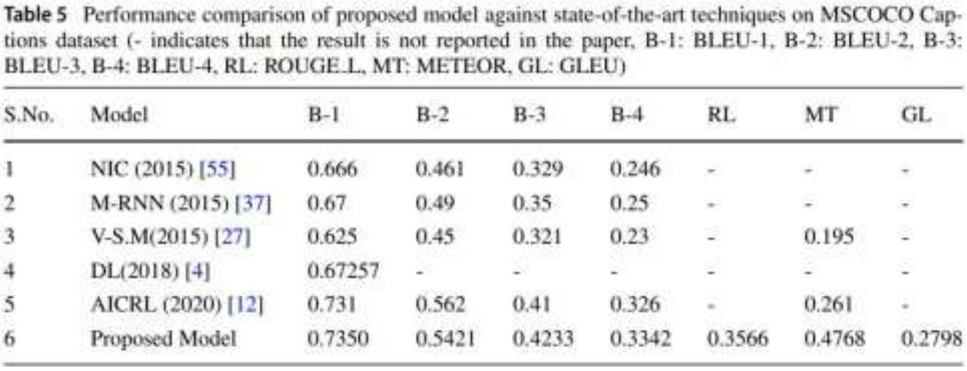


Table 5 Performance comparison of proposed model against state-of-the-art techniques on MSCOCO Captions dataset (- indicates that the result is not reported in the paper, B-1: BLEU-1, B-2: BLEU-2, B-3: BLEU-3, B-4: BLEU-4, RL: ROUGE_L, MT: METEOR, GL: GLEU)

S.No.	Model	B-1	B-2	B-3	B-4	RL	MT	GL
1	NIC (2015) [55]	0.666	0.461	0.329	0.246	-	-	-
2	M-RNN (2015) [37]	0.67	0.49	0.35	0.25	-	-	-
3	V-S.M(2015) [27]	0.625	0.45	0.321	0.23	-	0.195	-
4	DL(2018) [4]	0.67257	-	-	-	-	-	-
5	AICRL (2020) [12]	0.731	0.562	0.41	0.326	-	0.261	-
6	Proposed Model	0.7350	0.5421	0.4233	0.3342	0.3566	0.4768	0.2798

Throughout the experimentation, the following restrictions were noted. First of all, not all of the English grammatical rules were being followed by the generated captions. Second, there is room to improve the training process' efficiency given the high training time that was noted. Not to mention, compared to the reference captions, the generated captions seem less detailed.

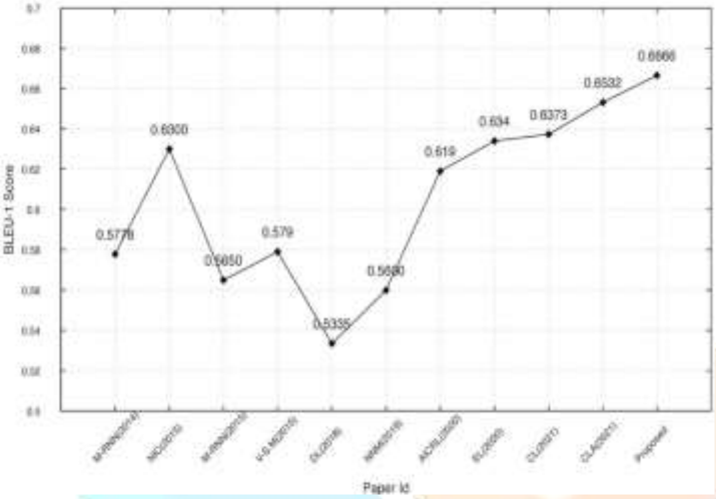


Fig. 5 Graph comparison of BLEU-1 scores of proposed models against state-of-the-art techniques on Flickr8k dataset. The values on the x-axis (Paper Id) are from 2nd column ("Model") of Table 4

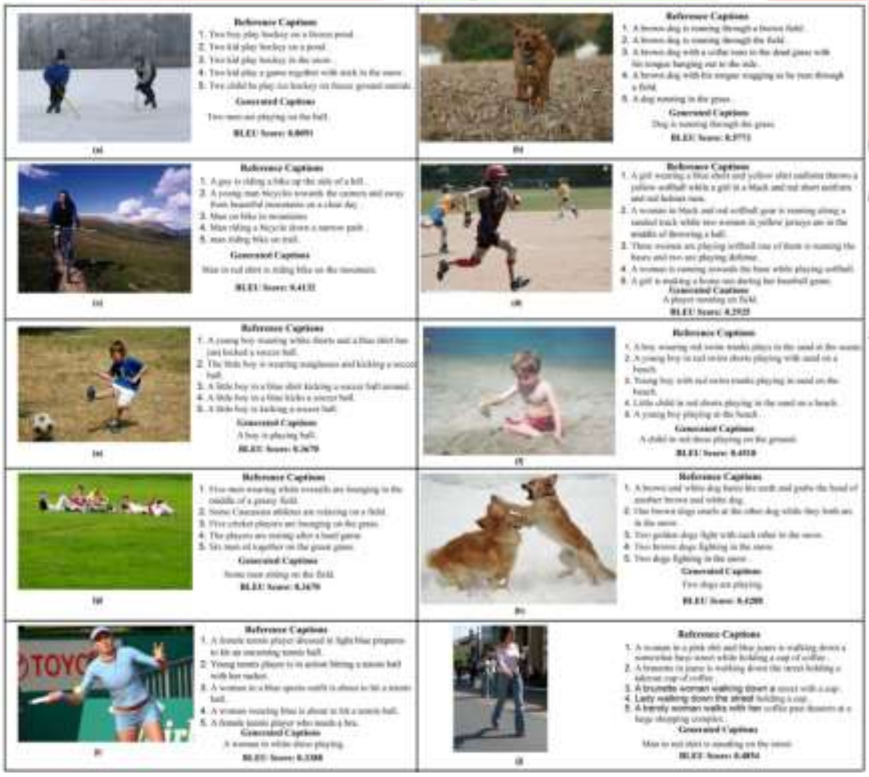


Fig. 6 Figure showing sample images with their reference captions along with generated caption using the proposed model

5 Conclusion and future work

Generating captions for images is a fascinating and challenging process with several applications in various domains, such as image organization and retrieval. An encoder-decoder based approach was suggested in this study to produce grammatically accurate image captions.

An LSTM-based decoder and a CNN-based encoder make up the suggested model. The model has been tested using Flickr8k and MS-COCO Captions datasets on a number of common metrics, including BLEU, METEOR, GLEU, and ROUGE L. Based on the experimental data, the model outperforms all current state-of-the-art methods in terms of BLEU, METEOR, and GLEU scores. The outcomes of caption generation on real sample photographs were also provided in the publication, supporting the validity of the suggested method. While B-LSTMs may need more tuning for optimal performance because to their increased number of parameters, LSTMs have been demonstrated to perform better.

Future methods may improve on this point, as the suggested model was extremely close to the best outcomes based on ROUGE L score. Additionally, experimental results may also be reported using other available metrics. Moreover, attention-based models could be used to increase the models' resilience. Additionally, alternative cross-domain strategies, such neuro-symbolic strategies, might be favored to improve the models' capacity for explanation as well as the logic behind coming up with certain descriptions for related photos [46]. Video inputs can also be captioned using image captioning, which can be used to identify significant events in a video over a period of time. This might have a significant impact on applications like home monitoring and surveillance.

References

1. Aggarwal AK (2022) Learning texture features from glcm for classification of brain tumor mri images using random forest classifier. Trans Signal Process 18:60–63
2. Albawi S, Mohammed TA, Al-Zawi S (2017) Understanding of a convolutional neural network. International Conference on Engineering and Technology (ICET), Antalya, Turkey, 2017, pp 1–6, https://doi.org/10.1109/ICEngTechnol.2017.830818
3. Ali Farhadi MH, Sadeghi MA, Young P, Rashtchian C, Hockenmaier J, Forsyth D (2010) Every picture tells a story: Generating sentences from images. In: European conference on computer vision, pp 15–29. Springer
4. Amritkar C, Jabade V (2018) Image caption generation using deep learning technique. In: Proceedings - 2018 4th International Conference on Computing, Communication control and automation ICCUBEA, 2018, pp 1–4
5. Arora K, Aggarwal AK (2018) Approaches for image database retrieval based on color, texture, and shape features. In: Handbook of research on advanced concepts in real-time image and video processing, pp 28–50. IGI Global
6. Bai S, An S (2018) A survey on automatic image caption generation. Neurocomputing 311:291–304
7. Barlas G, Veinidis C, Arampatzis A (2021) What we see in a photograph: content selection for image captioning. Vis Comput 37(6):1309–1326
8. Bayoudh K, Knani R, Hamdaoui F, Mtibaa A (2021) A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. The Visual Computer. pp 1–32
9. Biswas R, Barz M, Sonntag D (2020) Towards explanatory interactive image captioning using top-down and bottom-up features, beam search and re-ranking. KI-K ¨u,nstliche Intelligenz 34(4):571–584
10. Cao P, Yang Z, Sun L, Liang Y, Yang MQ, Guan R (2019) Image captioning with bidirectional semantic attention-based guiding of long short-term memory. Neural Process Lett 50(1):103–119
11. Chen H, Ding G, Lin Z, Guo Y, Shan C, Han J (2021) Image captioning with memorized knowledge. Cognit Comput 13(4):807–820
12. Chu Y, Yue X, Lei Y, Sergei M, Wang Z (2020) Automatic image captioning based on resnet50 and lstm with soft attention. Wireless Communications and Mobile Computing, pp 2020
13. Ding G, Chen M, Zhao S, Chen H, Han J, Liu Q (2019) Neural image caption generation with weighted training and reference. Cognit Comput 11(6):763–777
14. Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2017) Long-Term Recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Anal Mach Intell 39(4):677–691
15. Department of Computer Science Sulabh Katiyar, Borgohain SK, Silchar Engineering National Institute of Technology (2021) Comparative evaluation of cnn architectures for image caption generation. International Journal of Advanced Computer Science and Applications

16. Dong X, Long C, Xu W, Xiao C (2021) Dual graph convolutional networks with transformer and curriculum learning for image captioning. arXiv:2108.02366
17. Ghosh A, Dutta D, Moitra T (2020) A neural network framework to generate caption from images. Springer Nature Singapore Pte Ltd., pp 171–180
18. Gong Y, Wang L, Hodosh M, Hockenmaier J, Lazebnik S (2014) Improving image-sentence embeddings using large weakly annotated photo collections. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8692 LNCS(PART 4): 529–545
19. Graves A, Mohamed A, Hinton GE (2013) Speech recognition with deep recurrent neural networks. pp 6645–6649
20. Gupta N, Jalal AS (2020) Integration of textual cues for fine-grained image captioning using deep cnn and lstm. Neural Comput Applic 32(24):17899–17908
21. He C, Hu H (2019) Image captioning with text-based visual attention. Neural Process Lett 49(1):177–185
22. Hochreiter S, Schmidhuber ¨ J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
23. Hodosh M, Young P, Hockenmaier J (2015) Framing image description as a ranking task: Data, models and evaluation metrics. In: IJCAI International Joint Conference on Artificial Intelligence, 2015-Janua(Ijcai), pp 4188–4192
24. Hossain MD, Sohel F, Shiratuddin MF, Laga H (2018) A Comprehensive Survey of Deep Learning for Image Captioning. ACM Comput Surv Zakir Articl 0(0):36
25. Huang F, Li Z, Wei H, Zhang C, Ma H (2020) Boost image captioning with knowledge reasoning. Mach Learn 109(12):2313–2332
26. Jiang T, Zhang Z, Yang Y (2019) Modeling coverage with semantic embedding for image caption generation. Vis Comput 35(11):1655–1665
27. Karpathy A, Li F-F (2017) Deep visual-semantic alignments for generating image descriptions. IEEE Trans Pattern Anal Mach Intell 39(4):664–676
28. Katpally H, Bansal A (2020) Ensemble learning on deep neural networks for image caption generation. In: Proceedings- 14th IEEE International Conference on Semantic Computing, ICSC 2020, pp 61–68
29. Kaur A, Chauhan AS, Aggarwal AK (2022) Prediction of enhancers in dna sequence data using a hybrid cnn-dlstm model. IEEE/ACM Transactions on Computational Biology and Bioinformatics
30. Khan MJ, Curry E (2020) Neuro-symbolic visual reasoning for multimedia event processing: Overview, prospects and challenges. In: CIKM (Workshops)
31. Khan MJ, Khan MJ, Siddiqui AM, Khurshid K (2021) An automated and efficient convolutional architecture for disguise-invariant face recognition using noise-based data augmentation and deep transfer learning. The Visual Computer, pp 1–15