Original Article
ISSN (Online): 2582-7472

ShodhKosh: Journal of Visual and Performing Arts
June 2024 5(6), 473–481

A REVIEW ON AUTOMATIC IMAGE CAPTIONING GENERATION

Rachita Dubey[1], Rohit Miri[2]

[1] Computer Science & Engineering, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India
[2] Computer Science & Engineering, Dr. C. V. Raman University, Bilaspur, Chhattisgarh, India

ABSTRACT

Image caption is a very popular approach through which descriptive language can be generated in natural form. It is a difficult task in the field of artificial intelligence to assess an image and then write captions that are appropriate using computer vision approaches. The motive of the paper is to review the related studies in image captioning. Numerous studies on image captioning have been conducted, however optimum precision is still needed for accurate and precise captioning. To create well-organized sentences, a system that considers both semantic and syntactic factors is necessary. It is necessary to get the things that are over the image and explain how they link to one another or to express the activity in accordance with the situation in the image in order to get a better caption. The goal of image captioning can be accomplished using a variety of machine learning techniques, and numerous studies have used CNN, RNN, DNN, LSTM, and other approaches. The majority of researchers evaluated their systems using a variety of benchmarks, including Flickr8K, Flickr30K, MSCOCO and many more. However, Flickr8K, which has 8092 images or challenges to test the system's performance, is the most used dataset.

Keywords: Image Captioning, CNN, RNN, DNN, LSTM, MSCOCO, Flickr30k, Flickr8k

Corresponding Author
Rachita Dubey,
rachitadubey1991@gmail.com

DOI
10.29121/shodhkosh.v5.i6.2024.4050

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

1. INTRODUCTION

One of the ways to evaluate this world or to use machines to become more familiar with it is through the collection and analysis of visual data. Captioning for images is an analysis that transforms the images into understanding. In this field, there are two steps, the first of which involves the feature extraction from the image. The process of classifying or identifying items in an image based on their unique features or textures is also known as object identification or object classification. The system will be able to classify the object and determine what kinds of things are there in the image once the feature has been extracted. The system refers to information about the anticipated actions in the second phase based on object information and activity depicted in the image. The information that has been extracted in the form of keywords and afterwards combined to make a phrase that describes the image is known as the image caption. This is also referred to as semantic information. There isn't a specific region of interest (ROI) because the entire image is taken into account as the ROI, and the system is in charge of gathering all the data regarding the foreground and background of the image. There are various uses for image captions, and they are quite effective. The most common application is human computer interface, which makes it simpler to communicate and can be used to add subtitles to videos [1]. We

may give real-time updates about traffic emergencies, road conditions, and the surrounding area by using image captions and keeping safety in mind [2].

Figure 1. Image Captioning [2]

Encoders and decoders are both used in the natural language processing model for image captioning. There are so many different methods that may be used to do it. If a system is implemented using a conventional approach, it will not be able to produce precise captions because it is challenging and hard AI problem that calls for specific machine learning techniques using real-time data and allows for the extraction of more precise data with less processing time. The block diagram of the traditional model for captioning images is shown in Figure 2, where specific steps were required to obtain the caption for the image, such as feature extraction, encoder, a technique or method for analysis that produced text according to the image, and finally, the caption for the image was generated. The implementation of an image caption system can be done in one of two ways: either using a template-based method or an encoder-decoder-based method. Encoder-decoder methodology is more practical and effective because Convolutional Neural Network is involved in both stages, which is more ideal for extracting semantic information from the image.

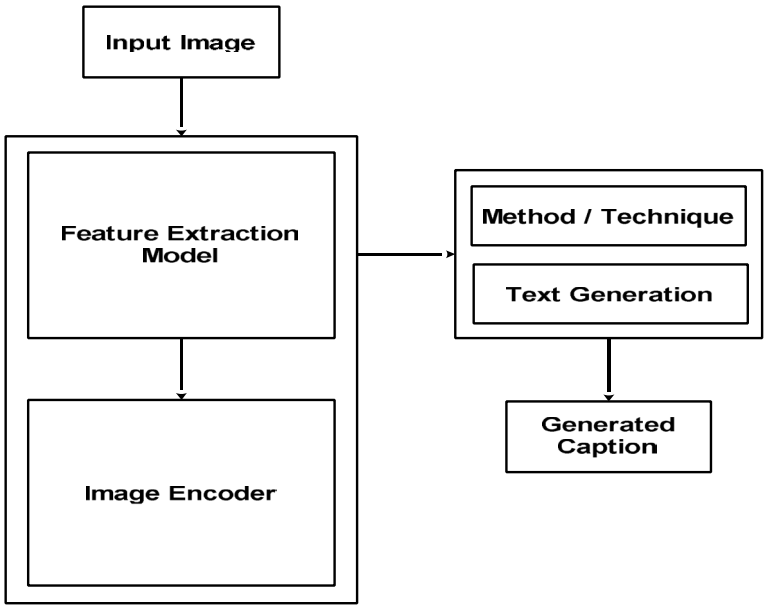

Figure 2. Conventional Approach for Image Captioning [2]

Template-based methodology is a weaker algorithm that is not suitable to forecast the precise output. When system combines image features along with semantic information then image caption can be generated.

2. RELATED WORK

This section discusses several studies that have been conducted in the area of creating picture captions, where different approaches have been employed to get exact information about the image. A stemming algorithm-related technique was introduced by K. Vijay et al. [4]. A natural language processing method called the stemming algorithm is used to extract the picture caption from various types of photos. It is a traditional strategy that has nothing to do with machine learning. It is founded on a morphological variation method that links words of a similar type. In order to get a higher level of accuracy with accurate captions, the strategy is template matching based, which is ineffective. A technique that is related to DenseNet was introduced by Xinru Wei et al. [5]. As the name implies, DenseNet is a very densely filled network with layers that have very high ranges, making the system more complex. As a result, the computation efficiency has suffered, and it needs a lot more memory to hold the trained weight model. The training phase takes very long time to complete. Due of its intricate layers, DenseNet is quite sluggish. The network should be smaller in size and employ effective pre-processing methods so that it may be assembled more quickly than networks like ResNet, VGG16, RNN, and many others. Niange Yu et al. [6] introduced a method that is related to Order Embedding algorithm. This type of algorithm aims to extract image attributes and keywords, then embed those keywords in accordance with the field-related topic. The keywords are organised according to the topic they cover. It includes the elements for the upper bound and lower bound. In order to get a higher level of precision and more precise captions, iteration has been used in both ways of the bidirectional technique, from top to bottom and bottom to top. Here, the system shows the results of the suggested approach using two datasets, MSCOCO and Flickr30K. They employed encoding and decoding techniques and chose a cutting-edge algorithm for comparison. Because it can be used with images of high quality, the embedding technique is a little subpar. A precise caption is important for high-quality images, but the object classification approach is compromised in some way by low pixel intensities in low resolution images.

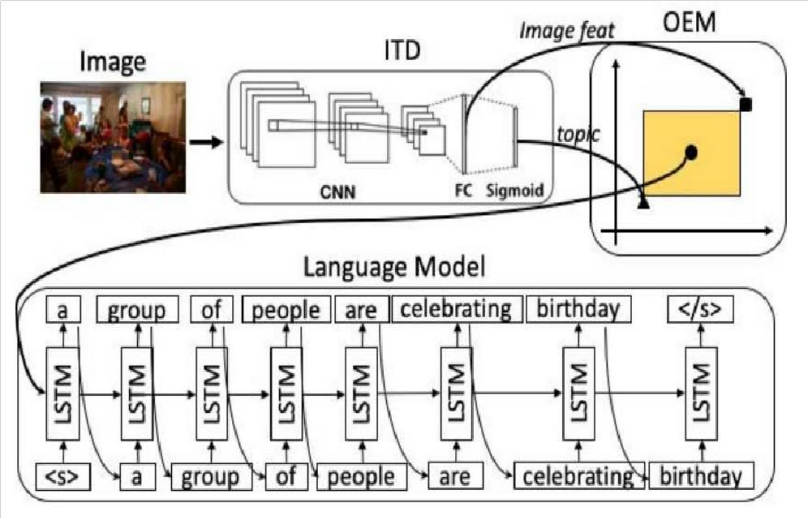

Figure 3. Order Embedding Model [6]

Min Yang et al. [7] introduced a method that is related to MLADIC which is a multitasking learning algorithm. For relating information from the source image to the target information, they used a two-phase algorithm. MSCOCO is used by the system to train the model, whereas Flickr30K and Oxford 102 are used for testing. Therefore, the first one is regarded as the source domain and the second one as the target domain. Through the use of a twofold learning mechanism, MLADIC carried out a number of exercises that simultaneously advance two connected goals in the hope that the execution of image captioning can be enhanced. The duty of creating literary textual information for a image is prepared with an encoder-decoder model. To include potential images based on text descriptions, the image amalgamation makes use of the constrained generative antagonistic network. In C-GAN, a discriminative model tries to identify the images while a generative model G orchestrates conceivable images given text representations. A two-venture technique is used to transfer information from the source spaces to the objective spaces in order to get around any obstacles that may exist between different locations. First, using the appropriate indicated source space information, the author pre-trained the model on the relationship between the network of images and text information. Image synthesis contains various

convolutional layers. Grishma Sharma et al. [8] introduced a method that is related to CNN and RNN models. As with the other models, they are utilising CNN in this instance to extract image characteristics for keyword analysis before sending the results to the RNN to further refine meaningful phrases. They stated that their model was superior to the GRU (Gated Recurrent Unit) model based on a comparison of their results. The system was also compared to the VGG16 model, and it was determined that CNN+RNN+LSTM could get superior results to the VGG16 models. RNNs are a subclass of artificial neural networks where the dynamic behaviour of the system is limited to the relationship between the nodes and coordinated graph along a brief succession. This enables it to display ephemeral data. RNNs can use their internal state (memory), similar to the feedforward neural network, to accommodate inputs of various lengths. They are hence suitable for undertakings. Recurrent neural networks are theoretically capable of managing arbitrary tasks to handle aggressive input successions. However, the same issue is still there, namely that RNN is a little slower than other models and that system training takes a lengthy time. Since it does not support vanished gradient, the vanished gradient can still be effectively used with this system or it can be elaborated. Ren C.Luo et al. [9] introduced a method that is related to Template augmentation method. They employed sentencebased templates, where image features were first extracted, followed by a comparison of the strings with the templates to determine the results. In an effort to preserve the relationship between the data, they tried to map the mixed data when training and testing the sets. The object recognition technique has been improved in order to optimise the system. In this study, the authors emphasise object identification or object categorization in order to identify the right object and produce the appropriate caption. However, the template-based approach is a highly traditional approach that excludes machine learning. The traditional approach uses pre-defined templates to produce results.

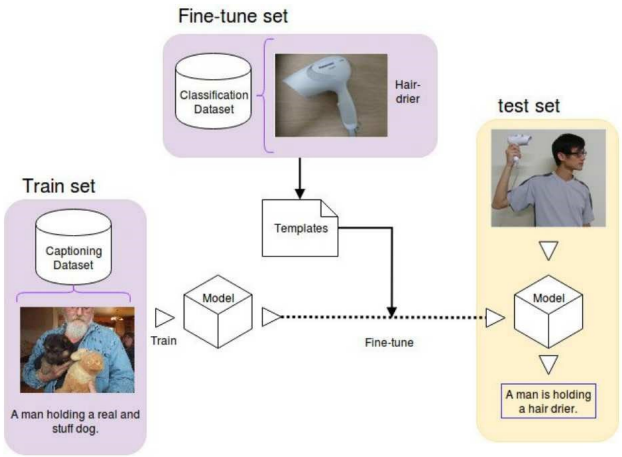

Figure 4. Template based Augmented Model [9]

The template-based augmented model for classifying the object and creating captions is shown in Fig. 4. The MSCOCO dataset, which contains 82783 photos and 5 descriptions for each of them, is used by the system to evaluate the model [9]. Tarun Wadhwa et al. [10] introduced a method that is related to CNN-RNN hybrid model for extracting the image features and LSTM for generating the image caption at last. The benchmark utilised by the authors, Flickr30K, has 31783 photos and 5 captions for each of them. To determine the model's accuracy, the output of the developed system is compared to captions created by humans. Here, the height, width, and edges of the objects in the photos are analysed using a CNN model, and the other characteristics are extracted using an RNN. It implies that the RNN model processes the output of the CNN before the LSTM is used to rearrange the retrieved keywords. The majority of systems use similar methodologies and claim to achieve different precision or scores by enhancing the system's training module or by employing other benchmarks. Depending on the needs of the system, the system uses a variety of layers, including convolutional, pooling, and maximum pooling layers. However, the system's accuracy fell short of expectations due to the vanishing problem and a large network. The system must overcome a number of difficulties when dealing with changed images or complex backgrounds where items appear to be somewhat smaller than they actually are. Even after thorough training, the system is unable to identify those objects, and its accuracy falls short of expectations. The spatial orientation of a picture is not supported by CNN, which performs slower and produces lower-quality results than maxpool. If an image has been manipulated, the system becomes confused and is unable to accurately characterise the image or produce a correct description in comparison to the values obtained from the real world.

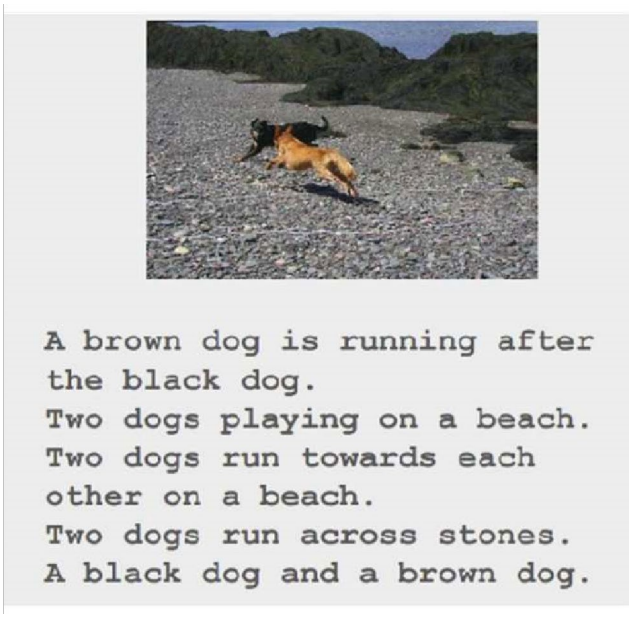

Figure 5. Model Output for Benchmark Flikr30k [10]

B.Krishnakumar et al. [11] introduced a method that is related to deep learning methodology. Additionally, they employed CNN and RNN learning models to analyse the image characteristics, and BLEU was used to calculate scores and compare them to actual values from the real world. They used datasets for picture captioning that contain 8091 photos, but the system was only trained on 6000 and evaluated on 1000, which is a bit fewer. It is necessary to test the system with at least 6000 photos in order to verify its legitimacy. Similar to the previous one, this one also suffers from the exploding gradient problem, and the network's complexity and slowness reduced the system's accuracy. Genc Hoxha et al. [12] introduced a method that is related to CNN (Convolutional Neural Network) and RNN (Recurrent Neural Network) training models. The beam searching algorithm used in this study produces several descriptions for a single image. Lexical analysis is used to examine the system's scores and compare them to the values obtained by ground truth. Beam search creates several captions and arranges them later to discover the best ones that are more appropriate for the image. However, the vanishing and exploding structure problem also exists in RNN. Compared to other machine learning techniques, the training is slow and the network topology is complex. Hidekazu Yanagimoto et al. [13] introduced a method that is related to attention mechanism. It is a multiple perception generator that regulates the flow of the captions. In the encoding phase, attention is a weight model, and LSTM is employed to decode the features. System employs VGG16 because it may be used to create various perceptual models and uses 16 layers to assess the system's performance. They employed 14 X 14 segment tensors and 128 bit dimension vectors to encode the input data. Because VGG16 is a hefty model, it takes longer to train the network. Due to the possibility of increased errors, it has a vanishing gradient problem, which lowers the system's efficiency. Small filters should be used by the network to make it speedier and take less calculation time.

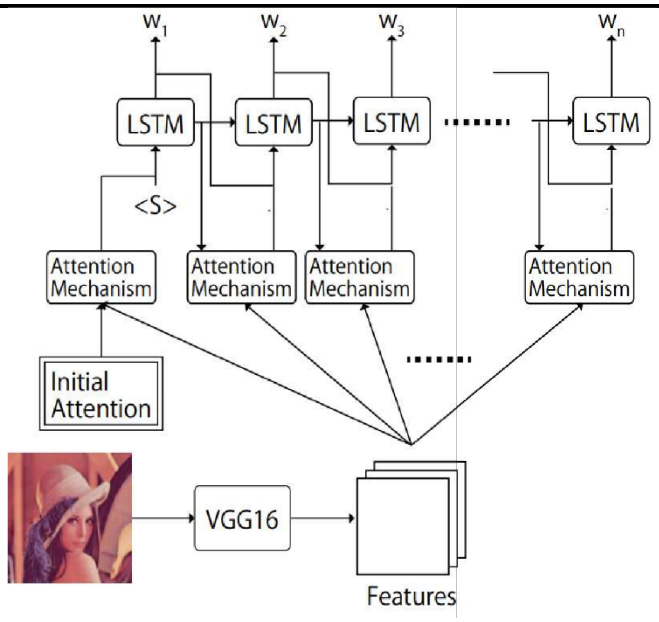

Figure 6. Attention Mechanism VGG16 Structure Model [13]

Figure 6 illustrates the structural model of the attention mechanism using the VGG16 model, which first extracts characteristics from the image before classifying the data using the LSTM model. The MSCOCO benchmark was employed by the authors to test their system, and they interpreted the results as such [13]. SiZhen Li et al. [14] introduced a method that is related to Deep learning algorithm. They created the caption using context coding. In this technique, the extracted features are extracted and combined using the SEnet-101 weight model. Additionally, they employed LSTM to create meaningful captions at the very end. By introducing more network layers, the authors of this article hope to improve the learning model's efficacy. As was already mentioned, LSTM has a flaw of its own that causes the system's accuracy to decline. Here, two separate scores of 0.783 and 1.176 have been achieved using BLEU1 and CIDER, respectively. The ResNet network is a deeper network, and one disadvantage of a deeper network is that it takes weeks to finish the system's training model.

Table 1. Comparison Matrices with certain algorithms

Algorithm [14]	B1	B4	METEOR
NIC	0.666	0.246	0.237
Hard-A	0.718	0.250	0.230
DeepVS	0.625	0.230	0.195
Adaptive	0.742	0.332	0.266
SCST	--------	0.313	0.260
MLO/ML	0.706	0.282	0.216
DL	0.783	0.361	0.264

Megha J Panicker et al. [15] introduced a method that is related to transfer learning algorithm that is related to the Convolutional Neural network (CNN). For more precision, this system combines CNN with Long Short Term Memory (LSTM). They conducted the testing phase using Flickr8K and obtained the results accordingly. Comparing the outcome to the extracted text is done using the Bilingual Evaluation Understudy. This system combined two different methods for creating captions to create a hybrid model. According on the retrieved keywords and the output of the image description, the system created scores that were compared to the values obtained from the real world. They trained with 6000 photos and evaluated 2000 photographs from the Flickr8K dataset. Because LSTM requires more time to train the model than other machine learning techniques, it has several downsides. Due to the extensive network, it also needs additional RAM. Implementing dropout is more challenging. The CNN and LSTM hybrid model is shown in Fig. 7 and has been used to do two tasks: first, an encoder to extract text, and second, a decoder to execute the final translation and create a coherent sentence.

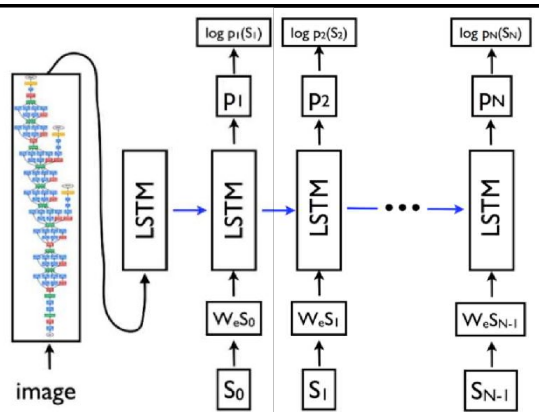

Figure 7. CNN & LSTM Structure Model [15]

Table 2. Comparison Matrices with certain algorithms

Method	Finding
CNN+ LSTM	It takes longer time to train the model as compared to the other machine learning approaches. It also requires more memory due to bulky network. Dropout is more complex to implement.
ResNet+ LSTM	It is a deeper network and one drawback of deeper network is that it takes weeks for completing the training model of the system.
VGG-16	VGG16 is a heavy weight model that requires more time for training the network. It has vanishing gradient problem due to that higher error may occur that degrade the system's performance. Network should be light in weight and use small filters for making it faster that requires less computation time.
CNN+ RNN	RNN has same vanishing and exploding structuring problem. The training is slow and the structure of the network is complex as compared to other machine learning approaches.
Template Augmentat-ion	Template based method is very conventional method where machine learning is not involved. Conventional method is returning the result on the basis of pre-defined templates.
Order Embedding	The embedding technique is bit poor because it can work with high quality images. High quality images pertain bit precise caption but poor resolution based images suffer somehow due to the low pixel intensities that degrade the object classification method.
DenseNet	DenseNet is very slow due to complex layers. Network should be lighter in weight and use proficient pre-processing techniques through which it can be compiled as faster network as compared to the ResNet, VGG-16, RNN and many more.

3. CONCLUSION

Automatic Image Caption Generation is a modern engineering technique that identifies objects from frames and captions the images in accordance with the scene depicted. Numerous studies have been examined in accordance with the paper's goals, and certain problems in some of them have been found. Numerous studies have been conducted using CN-RNN techniques. RNN, however, has a bit of a low calculation efficiency. The vanishing and exploding structural issue also affects RNN. In comparison to other machine learning techniques, the training is slow and the network structure is complex. Some studies use CNN+LSTM, which requires more time to train the model than the other machine learning techniques. Due to the extensive network, it also needs additional RAM. Implementing dropout is more challenging. Few are based on Order Embedding and Template Augmentation techniques. Methods based on templates and order embedding are fairly traditional methods that exclude machine learning. The traditional approach uses pre-defined templates to produce results. It is necessary to have a perfect system that can work with all types of datasets, achieve a high level of accuracy, and provide accurate captions. The network should be lightweight to enable a faster model in the system.

CONFLICTS OF INTEREST

None.

ACKNOWLEDGMENTS

None.

REFERENCES

Lee, S., & Kim, I. (2018). Multimodal feature learning for video captioning. Mathematical Problems in Engineering, 2018.

X. Wei, Y. Qi, J. Liu and F. Liu, "Image retrieval by dense caption reasoning," 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4, doi: 10.1109/VCIP.2017.8305157.

Pranoy Radhakrishnan, IIT Madras, Towards Data Science, (Sept 29, 2017).

K. Vijay and D. Ramya, "Generation of caption selection for news images using stemming algorithm," 2015 International Conference on Computation of Power, Energy, Information and Communication (ICCPEIC), 2015, pp. 0536-0540, doi: 10.1109/ICCPEIC.2015.7259513.

X. Wei, Y. Qi, J. Liu and F. Liu, "Image retrieval by dense caption reasoning," 2017 IEEE Visual Communications and Image Processing (VCIP), 2017, pp. 1-4, doi: 10.1109/VCIP.2017.8305157.

N. Yu, X. Hu, B. Song, J. Yang and J. Zhang, "Topic-Oriented Image Captioning Based on Order-Embedding," in IEEE Transactions on Image Processing, vol. 28, no. 6, pp. 2743-2754, June 2019, doi: 10.1109/TIP.2018.2889922.

M. Yang et al., "Multitask Learning for Cross-Domain Image Captioning," in IEEE Transactions on Multimedia, vol. 21, no. 4, pp. 1047-1061, April 2019, doi: 10.1109/TMM.2018.2869276.

Sharma, Grishma & Kalena, Priyanka & Malde, Nishi & Nair, Aromal & Parkar, Saurabh. (2019). Visual Image Caption Generator , Using Deep Learning. SSRN Electronic Journal. 10.2139/ssrn.3368837.

R. C. Luo, Y. -T. Hsu, Y. -C. Wen and H. -J. Ye, "Visual Image Caption Generation for Service Robotics and Industrial Applications," 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS), 2019, pp. 827-832, doi: 10.1109/ICPHYS.2019.8780171.

Tarun Wadhwa, Harleen Virk, Jagannath Aghav, Savita Borole Image. Captioning using Deep Learning. International Journal for Research in Applied Science & Engineering Technology (IJRASET) Volume 8 Issue VI June 2020.

B.Krishnakumar, K.Kousalya, S.Gokul, R.Karthikeyan, and D.Kaviyarasu, "IMAGE CAPTION GENERATOR USING DEEP LEARNING", (international Journal of Advanced Science and Technology- 2020)

G. Hoxha, F. Melgani and J. Slaghenauffi, "A New CNN-RNN Framework For Remote Sensing Image Captioning," 2020 Mediterranean and Middle-East Geoscience and Remote Sensing Symposium (M2GARSS), 2020, pp. 1-4, doi: 10.1109/M2GARSS47143.2020.9105191.

H. Yanagimoto and M. Shozu, "Multiple Perspective Caption Generation with Attention Mechanism," 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI), 2020, pp. 110-115, doi: 10.1109/IIAI-AAI50415.2020.00031.

S. Li and L. Huang, "Context-based Image Caption using Deep Learning," 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), 2021, pp. 820-823, doi: 10.1109/ICSP51882.2021.9408871.

Megha J Panicker, Vikas Upadhayay, Gunjan Sethi, Vrinda Mathur, Image Caption Generator, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-10 Issue-3, January 2021.

D.Kaviyarasu, B. , K. S. R. (2020). IMAGE CAPTION GENERATOR USING DEEP LEARNING. International Journal of Advanced Science and Technology, 29(3s), 975 - 980.

Aghav, Jagannath. (2020). Image Captioning using Deep Learning. International Journal for Research in Applied Science and Engineering Technology. 8. 1430-1435. 10.22214/ijraset.2020.6232.

Kesavan, Varsha & Muley, Vaidehi & Kolhekar, Megha. (2019). Deep Learning based Automatic Image Caption Generation. 1-6. 10.1109/GCAT47503.2019.8978293.

C. Amritkar and V. Jabade, "Image Caption Generation Using Deep Learning Technique," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-4, doi: 10.1109/ICCUBEA.2018.8697360.

S. M. Xi and Y. I. Cho, "Image caption automatic generation method based on weighted feature," 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), 2013, pp. 548-551, doi: 10.1109/ICCAS.2013.6703998.

Mathur, Pranay & Gill, Aman & Yadav, Aayush & Mishra, Anurag & Bansode, Nand. (2017). Camera2Caption: A real-time image caption generator. 1-6. 10.1109/ICCIDS.2017.8272660.
D. Elliott and F. Keller. Image description using visual dependency representations. In EMNLP, 2013.
A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young,C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In ECCV, 2010.
R. Gerber and H.-H. Nagel. Knowledge representation for the generation of quantified natural language descriptions of vehicle traffic in image sequences. In ICIP. IEEE, 1996.
Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In ECCV, 2014.
S. Hochreiter and J. Schmidhuber. Long short-term memory.Neural Computation, 9(8), 1997.
M. Hodosh, P. Young, and J. Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. JAIR, 47, 2013.
S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In arXiv:1502.03167, 2015.
A. Karpathy, A. Joulin, and L. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. NIPS,2014.
R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visualsemantic embeddings with multimodal neural language models. In arXiv:1411.2539, 2014.
R. Kiros and R. Z. R. Salakhutdinov. Multimodal neural language models. In NIPS Deep Learning Workshop, 2013.
G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Baby talk: Understanding and generating simple image descriptions. In CVPR, 2011.
P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi. Collective generation of natural image descriptions. In ACL, 2012.