Journal of Discrete Mathematical Sciences & Cryptography
ISSN 0972-0529 (Print), ISSN 2169-0065 (Online)
Vol. 26 (2023), No. 6, pp. 1801–1817
DOI : 10.47974/JDMSC-1728

Double awareness mechanism based deep learning framework for image captioning

Gaurav *
Pratistha Mathur †
Department of Information Technology
Manipal University Jaipur
Jaipur
Rajasthan
India

Abstract

Understanding the qualities of an image and converting them into a phrase or sentence that makes sense is the process of image captioning. Neuroscience research has only recently made clear the connection between human vision and language formation. Although there have been many methods for captioning images, including content retrieval and template filling, the current trend is toward deep learning-based methods. Using an image encoder, feature vectors are created from an image through the deep learning process, and a language decoder converts these feature vectors into a string of words. Using encoder-decoder approach or simple attention based has not provided so much efficient results. In the proposed model, double awareness-based mechanism has been used. The primary goal of this study is to extract visual properties from the region of interest (RoI) of an image as well as text features using the glove embedding technique. Inception ResNet version of Convolutional neural network (CNN) is used as an encoder. As a decoder, a gated recurrent unit is employed. The proposed model is tested on Flickr8k dataset and it can be seen that the results achieved through double awareness mechanism are highly effective.

Subject Classification: *68T07 Artificial neural networks and deep learning.*

Keywords: *Encoder, Decoder, Word embeddings, Image captioning.*

* *E-mail:* **gauravsingla31@gmail.com (Corresponding Author)**
† *E-mail:* **pratistha.mathur@jaipur.manipal.edu**

1. Introduction

Image captioning is a process of understanding the image features and representing these features into a meaningful phrase or sentence. Only in the last several years has neuroscience research revealed the relationship between human vision and language creation[1]. There have been various approaches to image captioning like content retrieval and template filling but current trend is towards the deep learning-based approaches. Through deep learning process, feature vectors are created from an image using an image encoder and these feature vectors are transformed into a sequence of words using language decoder[2]**.**

Variants of image captioning have been investigated as per the user needs and different captions styles. When the major focus is implicitly on low level visual features, image captioning can be perceptual. When the goal is to convey implicit and contextual information, it can be non-visual. However, when actual visual materials are presented, it can be conceptual [3].

The current trend to solve the task of image captioning is towards using the deep learning approach. In deep learning, Image encoder such as Convolutional Neural Network (CNN) is used to extract image features and Language decoder such as Long Short-Term Memory (LSTM) is used to convert these image features into an appropriate caption. [4]**.** Various deep learning-based approaches for image captioning have been presented in Figure 1. Recent improvements and innovations have been made in the usage of deep learning with relation to the attention mechanism, which concentrates on the most significant components of the input data. The following are the visual encoding strategies:

- Global CNN features (non-Attentive) [5]
- Region based features (Additive Attention) [6]
- Graphical attention [7]
- Self-attention [8]

A deep learning methodology uses language decoder to decode the image features received from image encoder. A language model's purpose is to predict the likelihood of a given sequence of words appearing in a phrase. A language model, or language decoder, of an algorithm for creating image captions assigns a probability to a sequence of n words as shown in (1):

$$P(Y_1, Y_2, Y_3, \ldots\ldots\ldots, Y_n) = \prod_{i=1}^{n} P(Y_i \mid Y_1, Y_2, Y_3, \ldots\ldots\ldots Y_{i-1}, X) \quad (1)$$

Where, X is representing the image encoding on which the language decoder is specifically conditioned. There are various language modeling strategies that has been used in past. The following are the major language modeling strategies [9]:

- Single or double layered LSTM based strategy [10].
- CNN based methods [11]
- Transformer based architectures [12]
- Image text early fusion based methods [13]

In the proposed model, double awareness-based mechanism has been used. The primary goal of this study is to extract visual properties from the region of interest (RoI) of an image as well as text features using the glove embedding technique. Inception ResNet version of Convolutional neural network (CNN) is used as an encoder[36-38]. As a decoder, a gated recurrent unit is employed. The proposed model is tested on Flickr8k dataset and it can be seen that the results achieved through double awareness mechanism are highly effective [39-40].

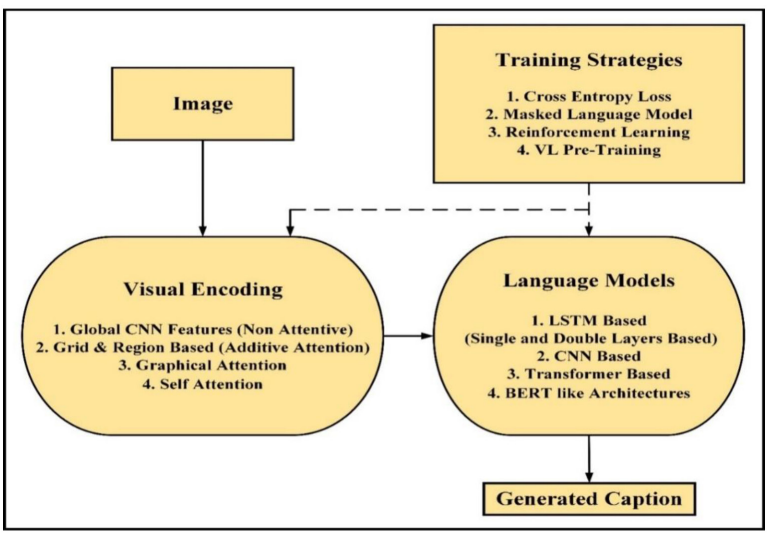

Figure 1

Deep Learning Based Approaches Developed for Image Captioning

2. Related Work

From simple encoder and decoder-based methods to attention-based models have been developed for image captioning. A model is proposed that employs a graph convolutional network to obtain the object-to-object spatial relationship inside the image and also uses a picture-level GCN to capture the distinctive information offered by related images. This process makes the framework better in understanding the relationship between objects and other related information[14][15]. The author uses a tool called the Global Enhanced Transformer, or GET. As part of GET, a Global Enhanced Encoder and a Global Adaptive Decoder are created for the purpose of incorporating a global feature and guiding caption generation, respectively[16][17][18]. One strategy proposed is based on two-fold strategy. Firstly, present a general multimodal model fusion framework for caption creation and Then, using the same fusion techniques, we combine a visual captioning model called Show, Attend, and Tell with BERT, to correct syntactic and semantic problems in captions [19]. An author suggests a brand-new, fully trainable deep learning image captioning model that extracts motion information from an image in order to provide captions. On two datasets, MSR-VTT2016-Image and a number of copyright-free photos, our suggested model was tested [20]. Based on one more study, the visual features of an image's region of interest (RoI) are retrieved and used as guiding data in gLSTM. This allows gLSTM to produce captions for images that are more accurate thanks to the visual elements of the RoI. It is suggested to use two approaches, one based on a region and the other on the complete image [21][22][23][24]. In one paper, a framework is put out that employs a gated recurrent unit as a decoder. Additionally, the suggested approach makes use of visual attention to enhance image attributes. Using the Flickr8K dataset, the proposed methodology has been put into practice [25]. A framework for image captioning called ATT-BM-SOM has been proposed in the following study. This model effectively fuses image data and creates high-quality captions, making up for the absence of image data selection and readable syntax [26][27][28]An Interactive Dual Generative Adversarial Network for picture captioning is suggested by the author of one publication. IDGAN has two generation-based generators and two retrieval-based generators, which mutually benefit from each other's complementing targets that are learned by two dual adversarial discriminators. To be more precise, the generation- and retrieval-based generators offer enhanced synthetic and recovered candidate captions with instructive feedback signals [29].

3. Proposed Model

The following stages outline some significant changes that have been made to the deep learning-based encoder and decoder technique in the proposed framework.

- Inception ResNetV2 is used as a convolutional neural network
- Region of Interest is identified for highly semantic visual features
- Glove embedding method is used to extract text based semantic features.
- A Gated Recurrent Unit serves as a decoder.

The encoder is used to extract an image's visual components. Visual properties from the region of interest (RoI) of an image are extracted using CNN and by applying visual enhancement techniques. Text features are extracted using the glove embedding technique. Inception ResNet version of Convolutional neural network (CNN) is used as an encoder. As a decoder, a gated recurrent unit is employed. The suggested method generates an output caption y for each input image that is a list of words encoded as 1-of-K vectors as shown in (2),

$$y = \{y_1, y_2, y_3, \ldots\ldots, y_n\},\ y_i \in R^K \tag{2}$$

where the captions can be as long as n characters, while the vocabulary can be as long as K characters. The pictorial representation of the proposed model is shown in Figure 2.

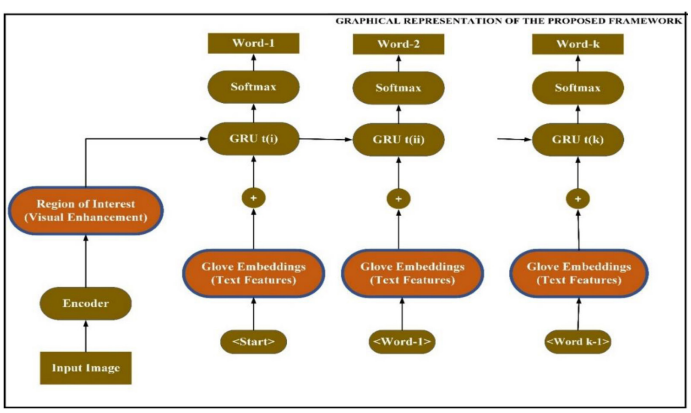

Figure 2

Double awareness deep learning based image captioning model

Initially a <start> word is taken and along with embedding information and region-based features from encoder are applied to GRU. The SoftMax activation function is applied onto the output of the GRU decoder. At time step 1, it will generate first word of the caption. At next time step, the output word generated is applied as an input word and the same process is applied again. Now, it will generate a new word for the caption after applying SoftMax activation function. This process is repeated until a specific word like <end> is produced by the decoder or the condition of number of words to be generated is fulfilled. At time step k, total k-1 number of words generated are applied as an input to the decoder.

3.1 *InceptionResNetV2*

With the help of Inception ResNet, features are taken out of the fully connected layer of the network, which creates the vector F that contains the overall information about the image as shown in (3)

$$F = Inception\_ResNet(I) \tag{3}$$

Here, F is a vector that is produced from the last convolutional layer of the CNN. This is fully connected layer and contains the overall information of the image.

3.2 *Gated Recurrent Unit (GRU)*

A form of RNN is LSTM. Because it addresses the problems with basic RNN, it is superior to simple RNN. Exploding gradient, disappearing gradient, and long-term dependency are two of the main problems that simple RNN encounters [30]. GRU is a more recent iteration of Recurrent Neural Network (RNN), which was created in 2014 and is rapidly gaining popularity. Traditional RNN has an issue with short-term memory. LSTM and GRU are two unique RNN subtypes. As GRU is lighter version of LSTM, the proposed model has used GRU as a decoder. GRU eliminates the RNN's gradient problem and is very beneficial for applications involving sentence production. Short-term memory and long-term memory are combined in GRU's hidden state. While the update gate of GRU understands how much of the past memory to preserve, the reset gate knows how much of the past memory to forget. GRU architecture is shown in Figure 3.

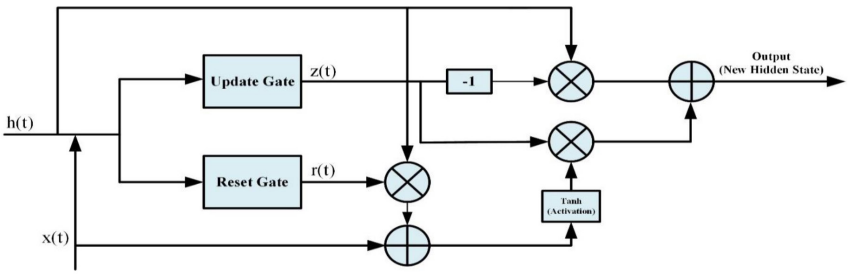

Figure 3

GRU Architecture

When captioning an image, GRU uses an input vector from CNN that represents theregion-based features of the image and word vectors using embedding methods then generates the caption for the one that has been provided. By using these data and an initial word, GRU predicts one word at a time and at next iteration this output word is used as an input word. The GRU works as in (4):

$$h_t = GRU(X_t, h_{t-1}) \tag{4}$$

Here X_t is the input vector and $h_{t-1}$is the previous hidden state. The internal mathematical equations of the GRU are as follows:

$$Z(t) = \sigma(W_Z.X_t + U_z.h_{t-1} + \beta_{update}) \tag{5}$$

$$R(t) = \sigma(W_R.X_t + U_r.h_{t-1} + \beta_{reset}) \tag{6}$$

$$h' = \tanh(W_h.X_t + R(t) * U_h.h_{t-1} + \beta_{update}) \tag{7}$$

$$h = Z(t) * h_{t-1} + (1 - Z(t)) * h' \tag{8}$$

Initially at (t-1) time step, pointwise multiplication takes place between update gate vector and hidden state, while the same operation is performed between input vector and the update gate weight vectors. The bias vector is multiplied with the guiding information vector and is combined with these two other vectors. At next time step (t), specified as in (5), the total result is subjected to the sigmoid function to produce a new update vector. One pointwise multiplication takes place between weight vector of reset gate and input vector. The multiplication of guiding information vector and bias vector is then added in addition to these two vectors. As demonstrated in (6), an activation function (sigmoid function) is applied to the overall outcome to produce a new vector for reset gateat every time step (t). The gate vector is updated and reset for time step (t).

A pointwise multiplication takes place between weight vector of hidden state and input vector. Pointwise multiplication of the update vector and regular multiplication of the recently created reset vector are applied to the concealed state vector. The outcome is included in the multiplication of guiding information vector and bias update vector. According to (7), it creates a somewhat new hidden state or output state. The update vector's negation is multiplied by this moderately created hidden state.Update vector is aggregated by the previous concealed state. As demonstrated in (8), these two different vectors are combined to produce the GRU's output at time step (t). The new hidden state is also the output at time step t.

3.3 *Visual Enhancement and Semantic Features Extraction*

Region based features are extracted by CNN using RoI segmentation. First, the difference of Gaussian (DoG) technique detects the initial important key points. The mathematical equation as in (9) for DoG is shown as:

$$F(x, y, \sigma, k) = M(x, y, \sigma k) - M(x, y, \sigma) \tag{9}$$

The image's pixel coordinates are (x,y), and the standard deviation of the normal distribution is σ. k is a multiplicative constant whose value is approximated to 1.4 for calculational purposes.Two-dimensional convolution function as in (10) for each image is calculate as:

$$M(x, y, \sigma) = H(x, y, \sigma) * i(x, y) \tag{10}$$

Where the calculation of G(x,y,σ) is shown as in (11):

$$H(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)} \tag{11}$$

Now, only dense points are saved and the rest are discarded using a filtering technique. Based on the number of key points, a threshold value is decided in between 1 and 5. Now statistic function is applied on each key point and output is compared with the threshold value. If output is larger or equal than the threshold value then it is kept as 1 otherwise it is kept as 0. Now, key points having value as 1 are dense points and are stored for further processing of finding the region of interest. The left and right boundaries are obtained using these dense key points and visual enhancement region is obtained as shown in Figure 4.

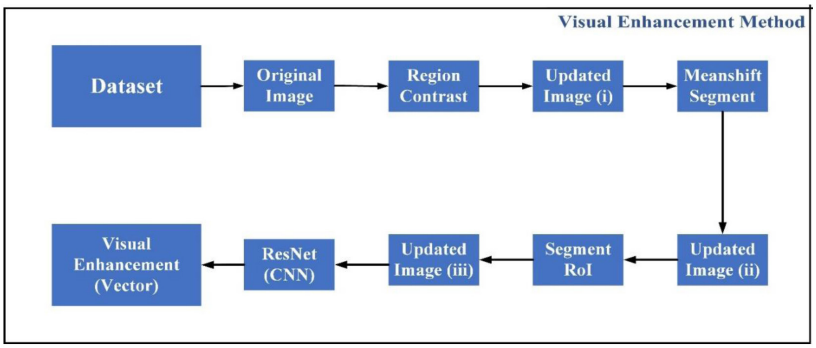

Figure 4

Visual Enhancement Method to Generate Region of Interest from an Image

3.4 *Word Embeddings*

In NLP systems, treating words as atomic units and representing them as vocabulary indexes is the most straightforward method of processing text. Word embeddings are a better method for generalization to cases that have not been observed before, because they may take into account the syntactic and semantic aspects of words. There are several word embedding techniques available, including GloVe, Word2Vec, and FastText[31]. Word embeddings are improved during the training process to increase the likelihood of other words occurring within a specific range in the graph. To find the optimum combinational architecture for a new application while taking contextual semantics into consideration, we have looked at and studied many images captioning models from a variety of areas[32]. Word embedding is important as a readable and grammatically sound sentence is essential for understanding and the order in which the words are used to create the captions is most significant [33]relatable meaning of the scenario inscribed. Different models can be used to accomplish this arduous task depending on the context and requirement of what needs to be achieved. An encoder–decoder model which uses the image feature vectors as an input to the encoder is often marked as one of the appropriate models to accomplish the captioning process. In the proposed work, a dual-modal transformer has been used which captures the intra-and inter-model interactions in a simultaneous manner within an attention block. The transformer architecture is quantitatively evaluated on a publicly available Microsoft Common Objects in Context (MS COCO. Glove embedding technique has been applied to the proposed model in

order to generate vector representations of the captions and extract text features.

4. Results

The most well-known image captioning datasets are MSCOCO, Flickr30K, and Flickr8K. There are five captions for each of the 8,000 photographs in the Flickr8K collection. In total, the Flickr8K dataset's photographs have 30.000 captions. The dataset, which include examples from numerous categories, are quite effective at captioning images. Images are selected from the databases and utilized for training, testing, and validation. The dataset used in this study for the model is Flickr8K. In Table 1, a summary of the Flickr8k dataset is displayed. 6,000 of the 8,000 total photos were used for training. The remaining 1,000 photos were utilized for validation and the first 1,000 were used for testing. However, the captions have been pre-processed so that words that might no longer be useful in creating better sentences for the search photos have been ignored.

For image captioning purposes, a variety of evaluation metrics are utilized. The model has been assessed in this study using the BLEU, METEOR, ROUGE, and CIDEr scores. The BLEU score measures how closely the generated and referenced captions match[34. The BLEU score ranges from 0 to 1. Scores of 0 indicate no matching between the generated and referred captions, whereas scores of 1 indicate complete word matching. BLEU-1, BLEU-2, BLEU-3, and BLEU-4 are the names of the four BLEU metrics scores. The BLEU-1 score demonstrates the correspondence of one word or grammatical unit between the generated and referred captions. Similarly, the other BLEU scores measure the correspondence of n words between referenced and generated captions. The main purpose of the METEOR evaluation measure is to concentrate

Table 1

Flickr8k dataset

	Dataset with Images	**Data size for Training**	**Data size for Testing**	**Data size for Development**
Flickr8K Dataset	8K Images	6K Images	1K Images	1K Images
	Captions for each image	**Dataset captions**	**Distinct words**	**Maximum caption length (In words)**
	5	40,000	8763	40

Table 2
Results of the experiment utilising different encoders as decoder

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
NIC (Inception V3) using LSTM	0.725	0.596	0.488	0.402	0.346	0.597	1.107
NIC (Inception V4) using LSTM	0.721	0.590	0.483	0.398	0.344	0.592	1.131
Proposed Model (Double awareness-based mechanism using ResNet and GRU)	0.752	0.621	0.518	0.429	0.382	0.615	1.119
An Attention Model (ResNet and GRU)	0.742	0.618	0.508	0.419	0.372	0.605	1.109

on word synonyms. The quality of the generated caption cannot be determined by the BLEU score alone. CIDEr and ROUGE are two other metrics that have been examined and are given in the final matrix[35]. The trial outcomes for the model using the Flickr8K Dataset are displayed in Table 2. According to the comparative study in Figure 5, the suggested model for image captioning has produced superior outcomes to a number of other current models. ResNet and Inception V3 encoders and GRU decoder along with double awareness mechanism are used to test the proposed model.

When compared to other existing models like NIC (InceptionV3 and LSTM), NIC (InceptionV4 and LSTM), and Attention Model (ResNet and GRU), the suggested model produced better results. One model employs LSTM as the decoder and InceptionV3 as the encoder. The BLEU scores (BLEU-1, BLEU-2, BLEU-3 and BLEU-4) in this model are 0.725, 0.596, 0.488, and 0.402, respectively.

The BLEU scores (BLEU-1, BLEU-2, BLEU-3 and BLEU-4) for ResNet and GRU based model, are 0.742, 0.618, 0.508, and 0.419, respectively. The proposed model, which combines a double awareness mechanism with a ResNet encoder and GRU decoder, has been found to produce the best outcomes of all these models. The BLEU scores (BLEU-1, BLEU-2, BLEU-3 and BLEU-4) for the proposed model are 0.752, 0.621, 0.518, and 0.429,

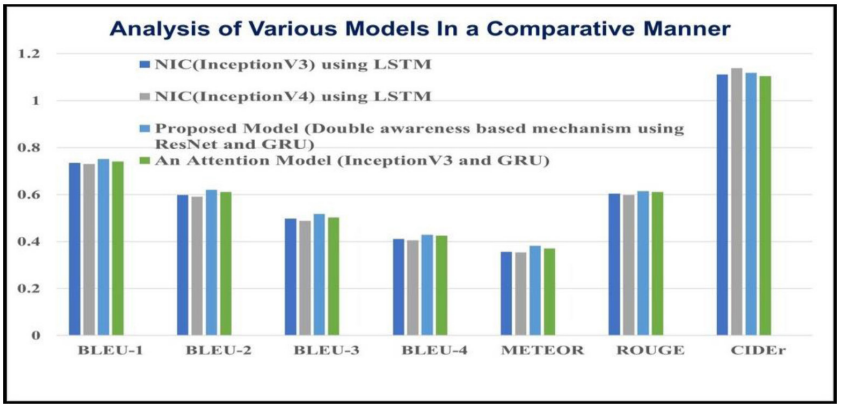

Figure 5

Analysis of Various Models in a comparative manner

respectively. In addition to BLEU scores, the suggested methodology has improved results for additional evaluation measures. On the Flickr8K dataset, the proposed approach is used to produce some samples of image captions in Figure 6. Sentences are grammatically and semantically sound, and generated captions are quite pertinent to the photographs.

Figure 6

Image caption creation utilising the suggested model

5. Conclusion

The encoder-and-decoder method, together with a double awareness-based mechanism, has been used in the study to implement image captioning. In the proposed model, double awareness-based mechanism has been used. The primay goal of this study is to extract visual properties from the region of interest (RoI) of an image as well as text features using the glove embedding technique. Inception ResNet version of Convolutional neural network (CNN) is used as an encoder. As a decoder, a gated recurrent unit is employed. The proposed model is tested on Flickr8k dataset and it can be seen that the results achieved through double awareness mechanism are highly effective. The suggested framework, outperformed rival frameworks in terms of BLEU scores. In the suggested paradigm, BLEU scores and other evaluation scores like CIDEr, METEOR, and ROUGE are much higher. Future study will put greater emphasis on using additional attention mechanisms and diverse datasets so that many more things can be covered for the purpose of image captioning.

References

[1] M. D. Zakir Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," ACM Comput. Surv., vol. 51, no. 6 (2019), doi: 10.1145/3295748.

[2] S. Kalra and A. Leekha, "Survey of convolutional neural networks for image captioning," *J. Inf. Optim. Sci.*, vol. 41, no. 1, pp. 239–260 (2020), doi: 10.1080/02522667.2020.1715602.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 07-12-June, pp. 3156–3164 (2015), doi: 10.1109/CVPR.2015.7298935.

[4] X. Liu, Q. Xu, and N. Wang, "A survey on deep neural network-based image captioning," Vis. Comput., vol. 35, no. 3, pp. 445–470 (2019), doi: 10.1007/s00371-018-1566-y.

[5] J. Wang, W. Wang, L. Wang, Z. Wang, D. D. Feng, and T. Tan, "Learning visual relationship and context-aware attention for image captioning," Pattern Recognit., vol. 98, p. 107075 (2020), doi: 10.1016/j.patcog.2019.107075.

[6] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep Hierarchical Encoder-Decoder Network for Image Captioning," IEEE Trans.

Multimed., vol. 21, no. 11, pp. 2942–2956 (2019), doi: 10.1109/TMM.2019.2915033.

[7] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, "Image caption generation with dual attention mechanism," Inf. Process. Manag., vol. 57, no. 2, p. 102178 (2020), doi: 10.1016/j.ipm.2019.102178.

[8] J. Liu, K. Cheng, H. Jin, and Z. Wu, "An Image Captioning Algorithm Based on Combination Attention Mechanism," Electron., vol. 11, no. 9 (2022), doi: 10.3390/electronics11091397.

[9] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 6077–6086 (2018), doi: 10.1109/CVPR.2018.00636.

[10] H. Sharma, M. Agrahari, S. K. Singh, M. Firoj, and R. K. Mishra, "Image Captioning: A Comprehensive Survey," 2020 Int. Conf. Power Electron. IoT Appl. Renew. Energy its Control. PARC 2020, pp. 325–328 (2020), doi: 10.1109/PARC49193.2020.236619.

[11] S. Katiyar and S. K. Borgohain, "Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation" (2021), [Online]. Available: http://arxiv.org/abs/2102.11237.

[12] Y. H. Chang, Y. J. Chen, R. H. Huang, and Y. T. Yu, "Enhanced image captioning with color recognition using deep learning methods," Appl. Sci., vol. 12, no. 1 (2022), doi: 10.3390/app12010209.

[13] J. H. Lim, C. S. Chan, K. W. Ng, L. Fan, and Q. Yang, "Protect, show, attend and tell: Empowering image captioning models with ownership protection," Pattern Recognit., vol. 122, p. 108285 (2022), doi: 10.1016/j.patcog.2021.108285.

[14] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual Graph Convolutional Networks with Transformer and Curriculum Learning for Image Captioning," MM 2021 - Proc. 29th ACM Int. Conf. Multimed., pp. 2615–2624 (2021), doi: 10.1145/3474085.3475439.

[15] E. Cetinic, "Iconographic Image Captioning for Artworks," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 12663 LNCS, pp. 502–516 (2021), doi: 10.1007/978-3-030-68796-0_36.

[16] J. Ji et al., "Improving Image Captioning by Leveraging Intra- and Inter-layer Global Representation in Transformer Network" (2020), [Online]. Available: http://arxiv.org/abs/2012.07061.

[17] S. Dubey, F. Olimov, M. A. Rafique, J. Kim, and M. Jeon, "Label-Attention Transformer with Geometrically Coherent Objects for Image Captioning," pp. 1–14 (2021), [Online]. Available: http://arxiv.org/abs/2109.07799.

[18] A. U. Haque, S. Ghani, and M. Saeed, "Image Captioning with Positional and Geometrical Semantics," IEEE Access, vol. 9, pp. 160917–160925 (2021), doi: 10.1109/ACCESS.2021.3131343.

[19] M. Kalimuthu, A. Mogadala, M. Mosbach, and D. Klakow, Fusion Models for Improved Image Captioning, vol. 12666 LNCS. Springer International Publishing (2021).

[20] K. Iwamura, J. Y. L. Kasahara, A. Moro, A. Yamashita, and H. Asama, "Potential of Incorporating Motion Estimation for Image Captioning," 2021 IEEE/SICE Int. Symp. Syst. Integr. SII 2021, pp. 23–28 (2021), doi: 10.1109/IEEECONF49454.2021.9382725.

[21] J. Zhang, K. Li, Z. Wang, X. Zhao, and Z. Wang, "Visual enhanced gLSTM for image captioning," Expert Syst. Appl., vol. 184, no. June, p. 115462 (2021), doi: 10.1016/j.eswa.2021.115462.

[22] I. Azhar, I. Afyouni, and A. Elnagar, "Facilitated deep learning models for image captioning," 2021 55th Annu. Conf. Inf. Sci. Syst. CISS 2021 (2021), doi: 10.1109/CISS50987.2021.9400209.

[23] B. Wan, W. Jiang, Y. M. Fang, M. Zhu, Q. Li, and Y. Liu, "Revisiting image captioning via maximum discrepancy competition," Pattern Recognit., vol. 122, p. 108358 (2022), doi: 10.1016/j.patcog.2021.108358.

[24] Gaurav and P. Mathur, "A Survey on Various Deep Learning Models for Automatic Image Captioning," *J. Phys. Conf. Ser.*, vol. 1950, no. 1 (2021), doi: 10.1088/1742-6596/1950/1/012045.

[25] Gaurav and P. Mathur, "An Attention Mechanism and GRU Based Deep Learning Model for Automatic Image Captioning," *Int. J. Eng. Trends Technol.*, vol. 70, no. 3, pp. 302–309 (2022), doi: 10.14445/22315381/IJETT-V70I3P234.

[26] Z. Yang and Q. Liu, "ATT-BM-SOM: A Framework of Effectively Choosing Image Information and Optimizing Syntax for Image Captioning," IEEE Access, vol. 8, pp. 50565–50573 (2020), doi: 10.1109/ACCESS.2020.2980578.

[27] Z. Deng, Z. Jiang, R. Lan, W. Huang, and X. Luo, "Image captioning using DenseNet network and adaptive attention," Signal Process.

Image Commun., vol. 85, p. 115836 (2020), doi: 10.1016/j.image.2020.115836.

[28] X. Zhang, S. He, X. Song, R. W. H. Lau, J. Jiao, and Q. Ye, "Image captioning via semantic element embedding," Neurocomputing, vol. 395, no. xxxx, pp. 212–221 (2020), doi: 10.1016/j.neucom.2018.02.112.

[29] J. Liu et al., "Interactive dual generative adversarial networks for image captioning," AAAI 2020 - 34th AAAI Conf. Artif. Intell., pp. 11588–11595 (2020), doi: 10.1609/aaai.v34i07.6826.

[30] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor, and M. Khan, "Image captioning using deep learning: A systematic literature review," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 5, pp. 278–286 (2020), doi: 10.14569/IJACSA.2020.0110537.

[31] D. H. Fudholi, A. Zahra, and R. A. N. Nayoan, "A Study on Visual Understanding Image Captioning using Different Word Embeddings and CNN-Based Feature Extractions," Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control, vol. 4, no. 1, pp. 91–98 (2022), doi: 10.22219/kinetik.v7i1.1394.

[32] U. Sirisha and B. Sai Chandana, "Semantic interdisciplinary evaluation of image captioning models," Cogent Eng., vol. 9, no. 1 (2022), doi: 10.1080/23311916.2022.2104333.

[33] D. Kumar, V. Srivastava, D. E. Popescu, and J. D. Hemanth, "Dual-Modal Transformer with Enhanced Inter-and Intra-Modality Interactions for Image Captioning," Appl. Sci., vol. 12, no. 13 (2022), doi: 10.3390/app12136733.

[34] P. Girdhar, P. Johri, and D. Virmani, "Incept_LSTM : Accession for human activity concession in automatic surveillance," *J. Discret. Math. Sci. Cryptogr.*, vol. 25, no. 8, pp. 2259–2273 (2022), doi: 10.1080/09720529.2020.1804132.

[35] M. Rajani Shree and B. R. Shambhavi, "POS tagger model for Kannada text with CRF++ and deep learning approaches," *J. Discret. Math. Sci. Cryptogr.*, vol. 23, no. 2, pp. 485–493 (2020), doi: 10.1080/09720529.2020.1728902.

[36] Kumar, A., Singh, K., & Khan, T., L-RTAM: Logarithm based reliable trust assessment model for WBSNs. *Journal of Discrete Mathematical Sciences and Cryptography*, 24(6), 1701-1716 (2021).

[37] Khan, T., & Singh, K., Resource management based secure trust model for WSN. *Journal of Discrete Mathematical Sciences and Cryptography,* 22(8), 1453-1462 (2019).

[38] Umamaheswaran, S., John, R., Deepthi, S. S., & Dharmarajlu, S. M., Caption positioning structure for hard of hearing people using deep learning method. *Journal of Discrete Mathematical Sciences and Cryptography*, 25(3), 623-633 (2022).

[39] Amrutha, K., & Prabu, P., Effortless and beneficial processing of natural languages using transformers. *Journal of Discrete Mathematical Sciences and Cryptography*, 25(7), 1987-2005 (2022).

[40] Hore, P., & Sharma, A., Code-switched end-to-end Marathi speech recognition for especially abled people. *Journal of Discrete Mathematical Sciences and Cryptography*, 25(3), 771-784 (2022).

Received February, 2023