International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V69I12P232
Volume 69 Issue 12, 267-279, December, 2021

An Intelligent Image Captioning Generator using Multi-Head Attention Transformer

Jansi Rani. J[1], Kirubagari. B[2]

[1]*Research Scholar,* [2] *Associate Professor, Department of Computer Science Engineering, Annamalai University, Chidambaram, India*

[1]jjansirani@yahoo.co.in, [2] kirubacdm@gmail.com

Abstract **-** *Recently, the advancements of artificial intelligence(AI) techniques have gained significant attention among research communications. At the same time, image captioning becomes an essential process in scene understanding, which involves the automated generation of natural language explanations dependent upon the content that exists in the image. The applicability of the image captioning process becomes important. With the development of deep learning (DL) and effective labeling datasets, image captioning approaches have been presented rapidly. In this aspect, this study designs an Intelligent Image Captioning Generator (IICG) model. The proposed IICG model technique encompasses different stages of preprocessing on image captions, namely, removal of punctuation marks, removal of single-letter characters, removal of numerals, and text vectorization. Besides, the DL-based DenseNet121 model is employed for the feature extraction process of the images. Then, the image captioning process takes place using the Multi-Head Attention Layer Transformer model, which consists of multiple encoders as well as decoders. The performance validation of the presented technique occurs utilizing Flickr 8k Dataset. A detailed comparative outcomes analysis is made, and the experimental outcomes demonstrate the superior performance of the proposed model in terms of Bilingual Evaluation Understudy (BLEU) score, ROUGE (Recall-Oriented Understudy for Gisting Evaluation), METEOR (Metric for Calculation of Translation with Explicit Ordering), CIDEr (Consensus-based Image Description Evaluation).*

Keywords **-** *Image captioning, Deep learning, Adam optimizer, DenseNet121, Flickr 8k dataset*

I. INTRODUCTION

The growth of the image description method might assist persons with physical disabilities to "see" the world. Currently, it has attracted growing interest and has become a hot research topic in the fields of computer vision (CV) [1]. Earlier image description generation systems were modeled using statistical language models and aggregating image data utilizing static object class libraries from the image. Instead of manual processes, it encompasses the automated analyzing, acquisition, and processing of big data for specific purposes to extract intuition and patterns. Generally, the CV model makes an effort to employ artificial intelligence (AI) algorithms, theory, frameworks, equations, and tools for completing the tasks of assisting computers to see as well as understand the contents of the analog and digital world via the imitation of the human visual scheme [2]. Even though understanding and seeing appear to be a very easy/trivial task for humans, it is nonetheless a challenging issue for computers partially due to our minimal understanding of how it processes things and how the human brain works. But, over many decades of investigation and advances in technology, some feats have been accomplished, and the CV model has been developing widely [3].

Nowadays, semantic segmentation remains a massive problem under the scope of video and image accepting along with image captioning that integrates the CV method with other fields of AI model termed Natural Language Process (NLP) for deriving sentence explanation of images [4]. Nevertheless, as with every remaining AI-based task, current subsets of machine learning (ML) methods, i.e., deep learning (DL) is the development of ML method, generating advanced outcomes in most of the tasks than conventional methods like Navie Bayes, Data Transformation Service, ensembles, clustering algorithms, and support vector machines (SVMs). The DL, as a domain of ML, employs a layer of artificial neural network for imitating the human neural network in decoder intuition from the huge number of information and is different from other ML approaches that are based largely on feature engineering, using area of interest under the form of feature extractor [5]. The stacked layers of Neural Networks signify feature hierarchy as a feature at the first layer is recreated from one 1layer to another in making difficult features [6]. Consequently, the deeper network is computationally intensive to train and model, which leads to the production of superior computational chips, involving Tensor Processing Unit (TPU) and Graphical Processing Unit (GPU).

Dependency method used for summarizing many web documents comprising data associated with image location and proposed a methodology for manually tagging geotagged images[7]. Kulkarni et al.presented an n-gram approach on the basis of network scale, collected candidate phrases, and merged them for make-up sentence-defining images from zero[8]. The language method is trained in the English Giga word corpus for attaining the estimate of movement under the images and likelihood of collocated scene, preposition, and nouns and employ this estimation as the parameter of the hidden Markov method[9]. The

image explanation can be attained using the prediction of more possible verbs, nouns, prepositions, and scenes which form the sentence. Kulkarni et al.proposed a detector for detecting objects from the image, classifying all the candidate regions and processing them through a prepositional relation function, and lastly apply CRF predictive image tags for generating natural language descriptions. Also, object detection has been implemented on the image[10].

This study designs an intelligent Image Captioning Generator (IICG) model. The proposed IICGmodel encompasses different stages of preprocessing on image preprocessing, captions preprocessing, namely removal of punctuation marks, removal of single-letter characters, removal of numerals, and text vectorization. Besides, the deep learning-based DenseNet121 model is employed for the feature extraction process of the images. Then, the image captioning process takes place using the Multi-Head Attention Layer Transformer model, which consists of multiple encoders as well as decoders. The performance validation of the proposed model takes place using Flickr 8k Dataset.

The paper is organized as follows: Section II explains Literature Review, Section III covers The Proposed IICG Model, Section IV explains Experimental Validation, Section V concludes the paper.

II. LITERATURE REVIEW

Chu et al. introduce one joint AICRL method that is capable of conducting the automated image captioning related to ResNet_50 as well as LSTM methods using soft attentions(SA) [11]. The AICRL method contains one decoder and one encoder. The decoder is developed using soft attention, LSTM, and RNN method, for selectively focusing on the attention on specific parts of the image for forecasting the following sentence. The encoder adapts ResNet50 based CNN method that generates wide representations of an image through embedded to set length vector. Lu et al. developed a fuzzy attention-based Dense Net, Bi-LSTM Chinese image captioning model. In the presented model, they enhance a densely connected framework for extracting features of an image at distinct scales and improve the model's capacity for capturing the weak features. Simultaneously, a Bi-LSTM is utilized as a decoder for enhancing the usage of contextual data[12]. The fuzzy attention method efficiently enhances the problem of correspondence among context information and image features.

In Zhang et al., the visual features of images region of interest (RoI) have been removed and employed as guiding data in gLSTM, where visual data of RoI is included asgLSTM for making precise image caption[13]. The two visual improved models depend on the region, and whole images are introduced correspondingly. Zhong et al. proposed a deep method for selecting significant sub-graphs and for decoding all elected subgraphs to an individual sentence. Through subgraphs, this method is capable of attending distinct modules of an image. Thus, the approach accounts for accurate, controllable, diverse, and grounded captioning simultaneously[14]. Liu et al. designed an automatic model to manifest construction activity scene through image captioning – a method rooted in natural language generation and CV [15]. The linguistic description scheme to manifest the scenes can be initially designed, and two exclusively devoted image captioning databases are generated for the validation system. Then, a common method of image captioning is introduced by integrating an encoder-decoder architecture using DNN models. Shen et al. developed a Variational Autoencoder, and Reinforcement Learning based Two-stage Multi-task Learning Model (VRTMM) for the remote sensing image captioning tasks [16]. Initially, finetune CNN together with VAE. Next, the Transformer generates the text description with semantic and spatial features. Then, the RL method is employed to enhance the quality of the sentences.

Del Chiaro et al. introduce an attention-based method that explicitly accommodates the transient nature of vocabularies from continuous image captioning tasks. That is, task vocabularies are not joint[17]. It is called Recurrent Attention to Transient Tasks (RATT), as well as shows how to adopt continuous learning algorithms-based weight regularization and knowledge distillation to a recurrent continuous learning problem. Guo et al. propose a new normalization approach and demonstrate that it is beneficial and possible for performing it on the hidden activation inside SA[18]. Next, compensate for the main limitation of the Transformer, which fails to construct the geometry form of the input object; also, they proposed a group of GSA models that expands SA to efficiently and explicitly consider the relative geometry connection among the objects in an image. For constructing the image captioning method, they integrate the two models and employ the vanilla self-attention network.

III. THE PROPOSED IICG MODEL

In this study, a new IICGmodel has been presented for effective image captioning. The proposed IICGmodel involves preprocessing in several ways, such as removal of punctuation marks, removal of single-letter characters, removal of numerals, and text vectorization. In addition, the features in the image are extracted by the DenseNet121 model, and the image captioning process is performed by the use of the Multi-Head Attention Layer. Fig. 1 demonstrates the overall block diagram of**IICG**model.

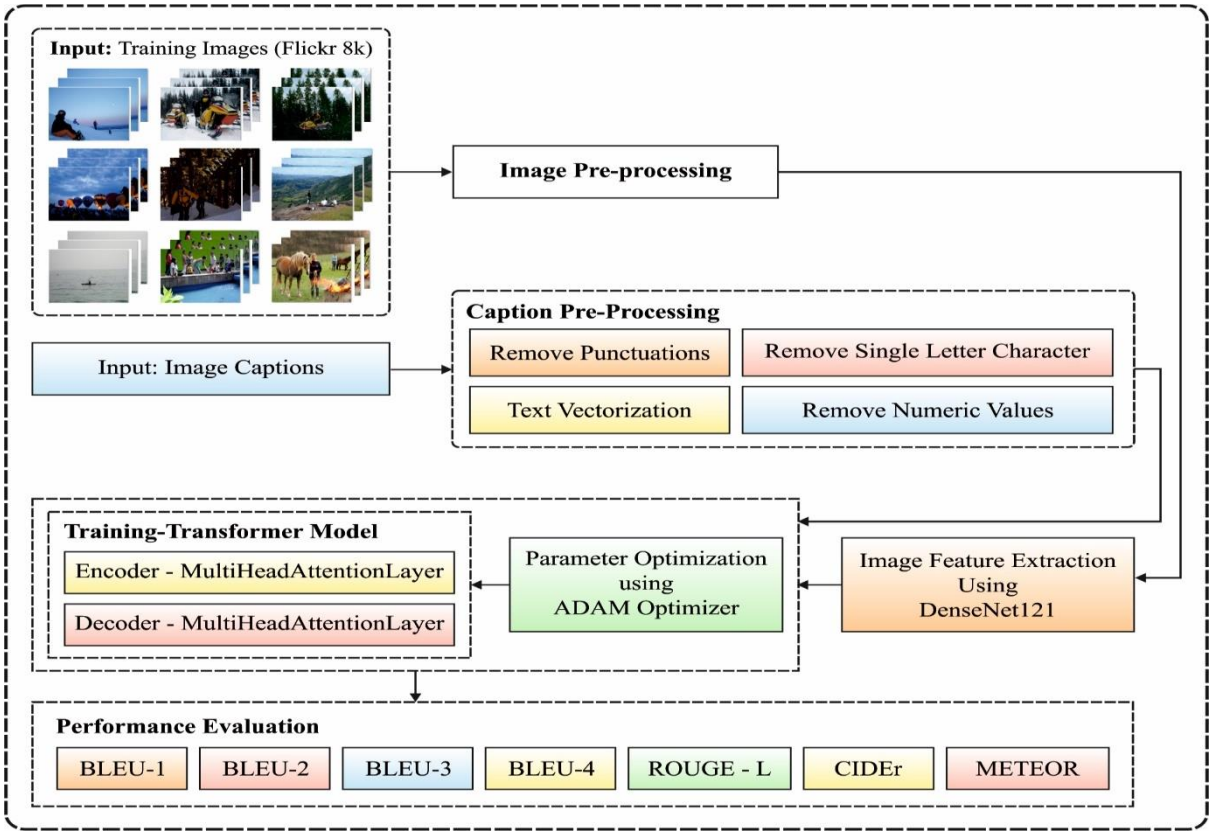

Fig. 1. Block diagram of proposed IICG model

A. Preprocessing

Primarily, the Images are pre-processed in the standard format, then text captions of the images are pre-processed for converting them to a useful format. Various sub-processes that exist in the preprocessing stage are listed as follows.

- Removal of single-letter words,
- Removal of multiple spaces,
- Removal of punctuation marks,
- Removal of numerals,
- Removal of stop words,
- Convert uppercase letters into lowercase, and
- Text vectorization using the Position Embedding approach.

B. Image Feature Extraction using DenseNet121 model

For extracting the features that exist in the images, the DenseNet121 model is utilized as transferring method. In a conventional feedforward CNN, all the convolution layers, excepting the first one (that take as an input), obtain the output of the prior convolution layers and generate an output feature map, i.e., later passed onto the following convolution layers. Thus, for the 'L' layer, it has a direct 'L' connection, one among all and the following layers. But, when the number of layers in the CNN increases, viz., when they get deeper, the 'gradient vanishing' problems emerge. It implies that the path for data from the input layer to the output layer rises; it could generate accurate data to get lost or 'vanish' that decreases the capacity of the network for efficiently training. DenseNet121 resolves these problems by altering the typical CNN framework and simplifying the connectivity patterns among the layers. In a DenseNet121 framework, every layer is directly connected to one another. Therefore it is termed Densely Connected Convolution Networks. For the 'L' layer, it has a direct L(L+1)/2 connection. The DenseNet121 component consists of:

- Connectivity
- Dense Block
- Growth Rate
- Bottleneck layer

In all the layers, the feature map of each prior layer isn't calculated but used and concatenated as an input. Accordingly, DenseNet requires fewer parameters when compared to conventional CNN. Also, it enables feature reuse as a redundant feature map is discarded. The usage of the concatenation process isn't possible once the size of the feature map get changes. But, an important role of CNN is the down-sampling of layers that decreases the size of feature maps via reduction dimension for gaining high computational speed. To allow this, Dense Net is separated into Dense Blocks, in which the dimension of the feature map has not been changed within a block. However, the number of filters among themselves is altered. Fig. 2 illustrates the architecture of the DenseNet-121 model.

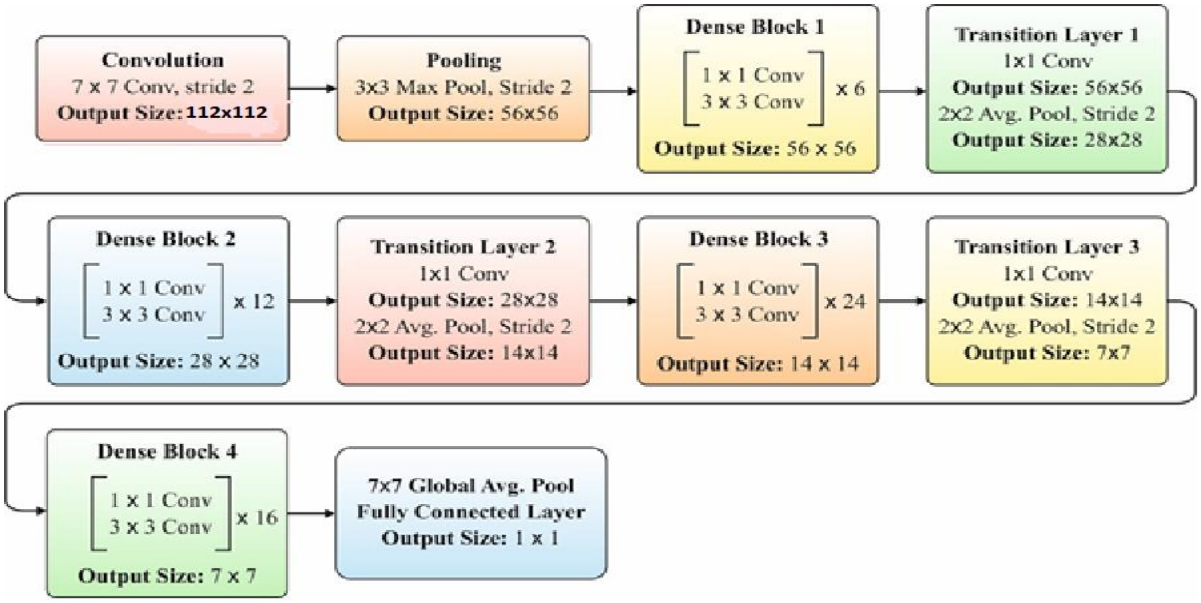

Fig. 2. Structure of DenseNet-121

The layers among the blocks are named Transition Layers that reduce large amounts of channels to half this of the present channel. For all the layers, Hl is determined as the composite function that employs 3 sequential operations: a convolution (Conv), batch normalization (BN), and ReLU activation. The size of the feature map raises when it passes through all the dense layers using 'K' features above the global state (present feature). The 'K' parameter is represented as the growth rate of networks that regulate the number of data included in all the layers [19]. When every H l function generates a k feature map, then the first layer contains

$$k_l = k_0 + k * (l - 1) \tag{1}$$

input feature map, in which k0 represents the number of channels in the input layer. Different from the present network architecture, DenseNet might contain a very narrow layer. Even though the layer only generates the k output feature map, the number of inputs could be relatively higher, particularly for additional layers. Therefore, a 1x1 convolutional layer could be presented as a bottleneck layer beforehand 3x3 convolutional layer for improving the speed and of efficiency the computations.

Briefly, DenseNet-121 contains fourAvgPool and 120 Convolutions. Because, transition layer outputs several redundant features, the layer in the 2nd and 3rd dense blocks allocate the minimum weight to the output of the transition layer. Although the weight of the full dense blocks is utilized by the last layer still, there might be higher-level features generated deeper to the module since they seem to be a high concentration towards the last feature map.

C. IICG Model Tuned using Adam Optimizer

For optimally adjusting the parameters involved in the DenseNet121 model, the Adam optimizer is utilized. During the Adam algorithm [20], the exponential decomposing average of the previous gradient m_k and previous squared gradient v_k can be expressed by:

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1) g_k, \tag{2}$$

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_k^2, \tag{3}$$

whereas g_k represent the gradient, $\beta_1 \& \beta_2$ denotes the decay rates that are closer to one. Note that $m_k \& v_k$ Our estimate of the initial moment (mean) and the 2nd moments (uncentered variance) of gradient correspondingly. This bias is countered by means of bias-corrected initial and 2nd-moment estimates, as follows

$$\hat{m}_k = \frac{m_k}{1 - \beta_1^k}, \tag{4}$$

$$\hat{v}_k = \frac{v_k}{1 - \beta_2^k}. \tag{5}$$

Therefore, Adam upgrading rule was given by:

$$w^{(k+1)} = w^{(k)} - \cdot \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \delta}, \tag{6}$$

Let δ be the smoothing term utilized for avoiding division by zero.

ALGORITHM 1: Adam algorithm

Data: assumed the first value $w^{(0)} = w_0$, the step size α, the toleranceε, the number of samples n and. Set $k = 0$.
Compute the augmented objective functions.
Evaluate the stochastic gradient.
Fixed the random index j.
Calculate the decaying average of a previous and previous squared gradient in Eqs. (2) and (3).
Compute the bias-corrected initial- and second-moment estimate in Eqs. (4) and (5).
Upgrade the vector $w^{(k)}$ From (6). If $\|w^{(k+1)} - w^{(k)}\| < \varepsilon$, then end the process. Or else, set $k = k + 1$, and repeat in Step 1.
Remark:
The default value for the decay rates is $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and the smoothing term is $\delta = 10^{-8}$, when the learning rate is $\alpha = 0.001$, and the tolerance is $\varepsilon = 10^{-6}$.

D. Image Captioning Process

Finally, the textual and images features derived in the previous stages are fed into the encoder unit in the transformer, which executes the actual image captioning process. Conventional image caption method-based visual attention model has been tried to advance in this study; hence an image caption mechanism-based multi-head attention model has been projected. It is made up of 6 identical sub-models, each containing 3 sublayers: multi-head spatial attention layer, full connection feed-forward layer, and multi-head self-attention layer. Amongst others, the extracting image feature from the CNN method is utilized and selected through the spatial attention model in the multi-head spatial attention sublayer. The syntactic feature in natural sentences is utilized and learned by the self-attention model in the multi-head self-attention sublayer. Mainly, the feature is combined with a full connection network layer.

The attention mechanism is an important portion of the compelling sequence modeling and transduction approach in image captioning, permitting modeling of dependency with no regard to their distance in the input or output series. The transformer network applies an encoder-decoder architecture as same as RNN. The architecture of the transformer [21] is given in Fig. 3. The major variation is that transformers can accept the input sentence/sequence in parallel, i.e., there is no time step related to the input, and every word in the sentence can be passed concurrently.

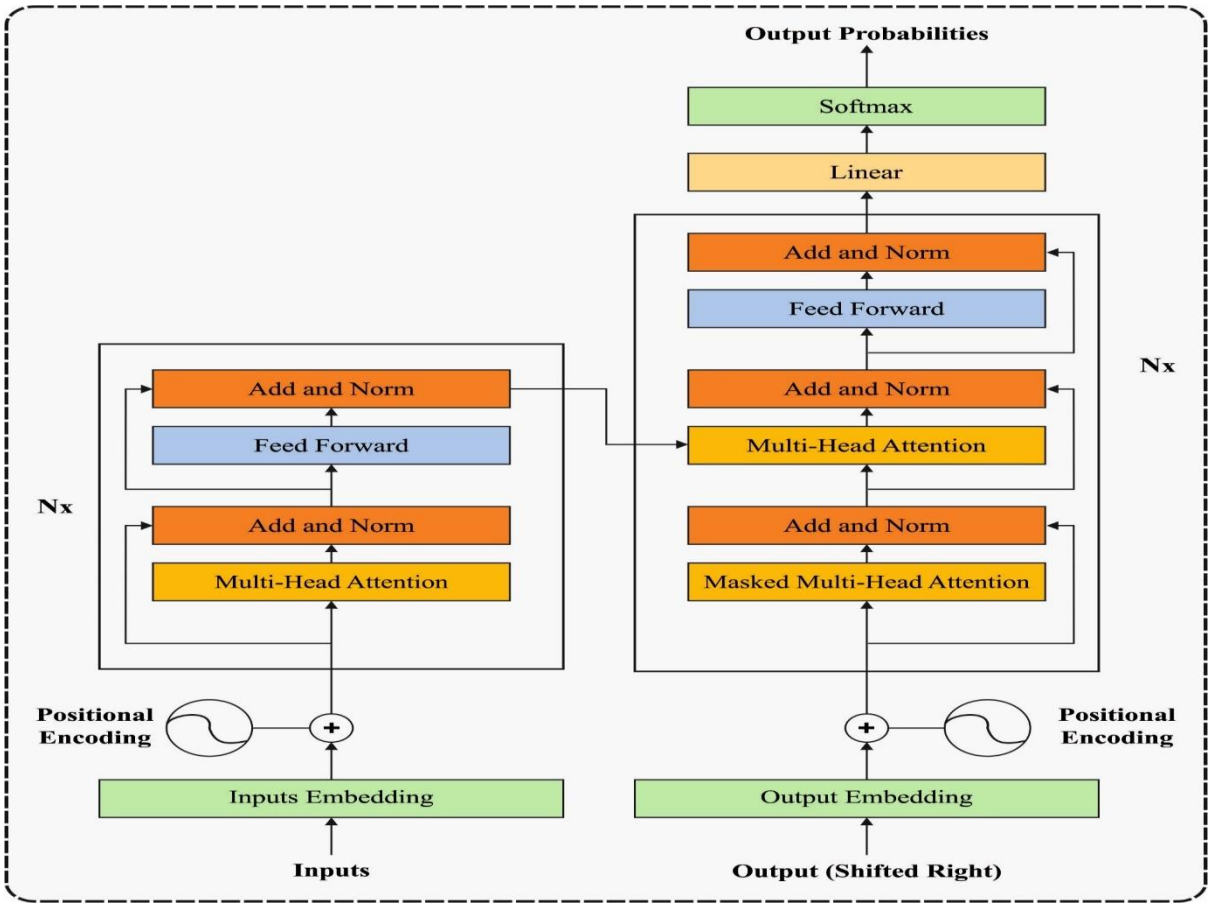

Fig. 3. Architecture of Transformer

a) Encoder Module

CNN is largely utilized for extracting image features in the Encoder model. The image feature would be extracted by DenseNets121 in the Encoder model. Hence the conventional DenseNet module couldn't be directly utilized since it is widely employed to resolve image classification. Rather than, the original DenseNet should be altered: eliminate the full connection layer, i.e., applied to classification, maintain the partial network to extract image features.

Based on this, there are 5 classes of convolution layers in DenseNet. Initially, the convolution layer includes a sequence of convolution layers that is generated from a dimension of 64 and size of 7x7 convolution kernels. Each 4-convolution layer consists of a convolution operation that is generated from a size of 1x1 convolution kernel. The difference for these 4 convolution layers is that the 2nd groups have 9 convolution layers; the 3rd groups contain 12 convolution operations; the 4th groups include 18 convolution layers; the final group is made up of 9 convolution layers.

b) Decoder Model

As there are 6 submodules in the entire Encoder model, all the sub-modules are comprised of a full connection feedforward layer, self-attention layer, and a spatial attention layer, and all the sublayers are made up of a level normalization operation and a residual connection [22]. Image features are extracted from distinct angles by using a distinct convolution kernel in the CNN method. When they learn this model for calculating attention through distinct weight parameter matrices several times rather than only once, and the multi-head attention is attained for every computing result, matrices are stitched together. The multi-head attention could learn from data related in distinct representation subspace, could also focus on distinct representation subspaces data from distinct locations. They could easily understand multi-head attention is that attention would be dot product for multiple times once the parameter isn't shared.

Hence we get a common equation for multi-head attention by:

$$head_i = Attention\left(QW_i^Q, QW_i^K, VW_i^V\right) \quad (7)$$

$$MultiHead(Q,K,V) = contact(head_1, \dots, head)W^{\circ} \quad (8)$$

So, we get the multi-head spatial attention model (Eq.(9)) by computing several spatial attention, in which X_e refers to the image feature vector matrix in the Encoder model, and X_d Denotes the d-dimension word vector matrix in the Decoder model. Hence we can get a multi-head self-attention model.

$$Y = MultiHead(Q,K,V) = MultiHead(X_d, X_e, X_e) \quad (9)$$

All the sub-modules include a full connection feedforward network in the Decoder model, has been shown in Eq. (10). The feedforward network contains two operations which contain two linear transformations and another container of one nonlinear transformation. Dropout is included among the two layers of the network to avoid overfitting.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (10)$$

Some absolute/relative positional data regarding vocabulary in the sequence needs to be included in the module for utilizing word sequence data. We get a positional coding data vector through a certain coding to code every word position when all the words have a certain position. A specific position data could be presented for every word when position embedding is integrated into a word embedding. Thus the word at distinct positions could be differentiated by the self-attention model. They employ sine and pre-function of distinct frequencies for position encoding:

$$PE(pos, 2i) = sin(pos/10000^{2i/d_{model}}) \quad (11)$$

$$PE(pos, 2i+1) = cos(pos/10000^{2i/d_{model}}) \quad (12)$$

Amongst others, pos denotes the position of the word in the sequence, as well, as i represent the ith component of the position-coding vector. Thus, the word could be mapped to the positional vector of the d_{model} using Eqs. (11) and (12). The decoder generates the text captions for the applied input images.

IV. EXPERIMENTAL VALIDATION

The proposed technique was simulated utilizing Python 3.6.5 tool. The parameter settings involved in the simulation is given as follows: IMAGE_SIZE = (224, 224), VOCAB_SIZE = 10000, SEQ_LENGTH = 20, EMBED_DIM = 512, NUM_HEADS = 2, FF_DIM = 512, BATCH_SIZE = 2, EPOCHS = 900, AUTOTUNE = tf.data.AUTOTUNE, and LEARNING_RATE = 0.001. The proposed model is tested using the Flickr8k dataset [23]. It consists of 8,000 images which are related to 5 distinct captions. Fig. 4 showcases a few sample images. In this study, differentassessment measures are used, which are given in the following.

- BLEU (BiLingual Evaluation Understudy): as metrics, it calculates the number of equivalent n-grams from the model predictions than that of ground truth.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): It can be helpful for summary estimation and is evaluated as an overlapping of one gram or bigrams among the predicted sequence and reference captions.
- METEOR (Metric for Calculation of Translation with Explicit Ordering): It tackles the drawbacks

of BLEU, as well as it depends on weighted F-score calculation and a penalty function intended for checking the direction of candidate sequence.

- CIDEr (Consensus-based Image Description Evaluation): It establishes the consensus among the predicted and reference sequences through TF-IDF weighting, stemming, and cosine similarity.

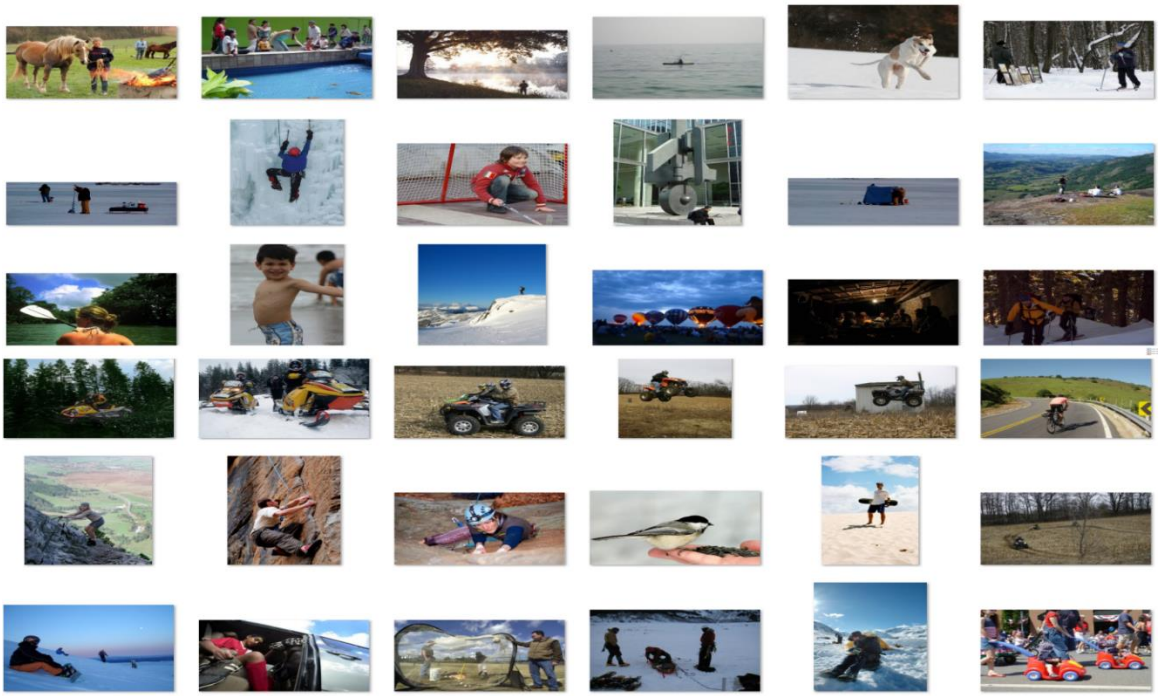

Fig. 4. Sample Images

Fig. 5 shows the sample results obtained by the proposed manner on the applied test images. The figures demonstrated that the proposed model has clearly captioned the images intelligently.

Fig. 5. Sample test images of IICG model

Fig. 6 demonstrates the accuracy analysis of the IICG model manner on the test Flickr8k dataset. The outcomes exhibited that the IICG model system has accomplished improved efficiency with enhanced training and validation accuracy. It can be observed that the IICGmodel manner has reached increased validation accuracy over the training accuracy.

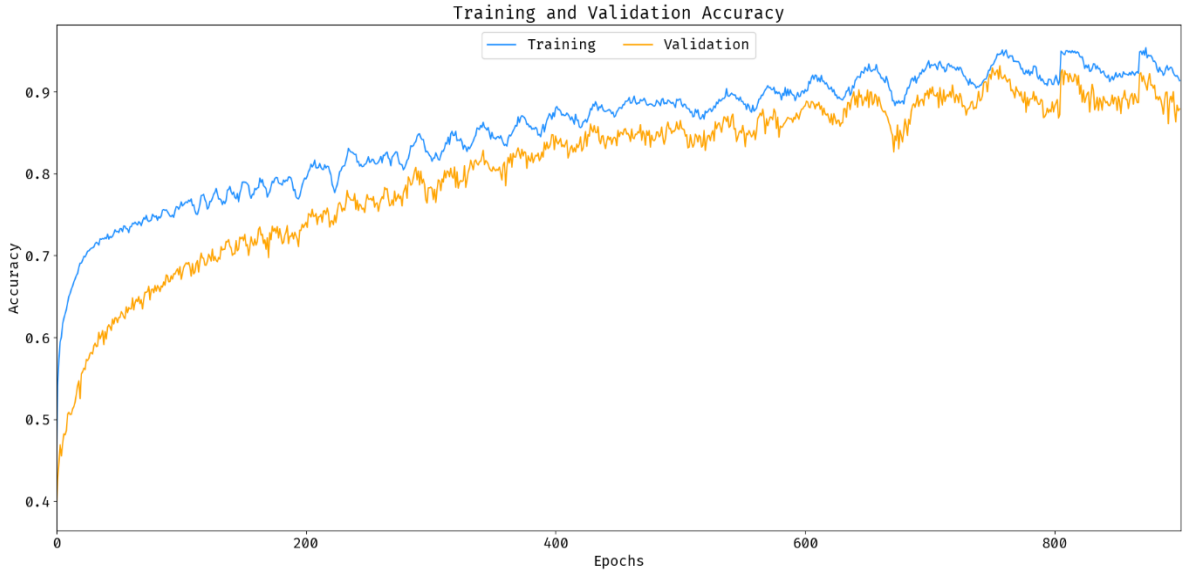

Fig. 6. Accuracy analysis of IICG model

Fig. 7 defines the loss analysis of the IICGmodel on the test Flickr8k dataset. The results portrayed that the IICGmodelapproach has resulted in a proficient outcome with the decreased training and validation loss. It can be experiential that the IICGmodel has accessible lower validation loss over the training loss.

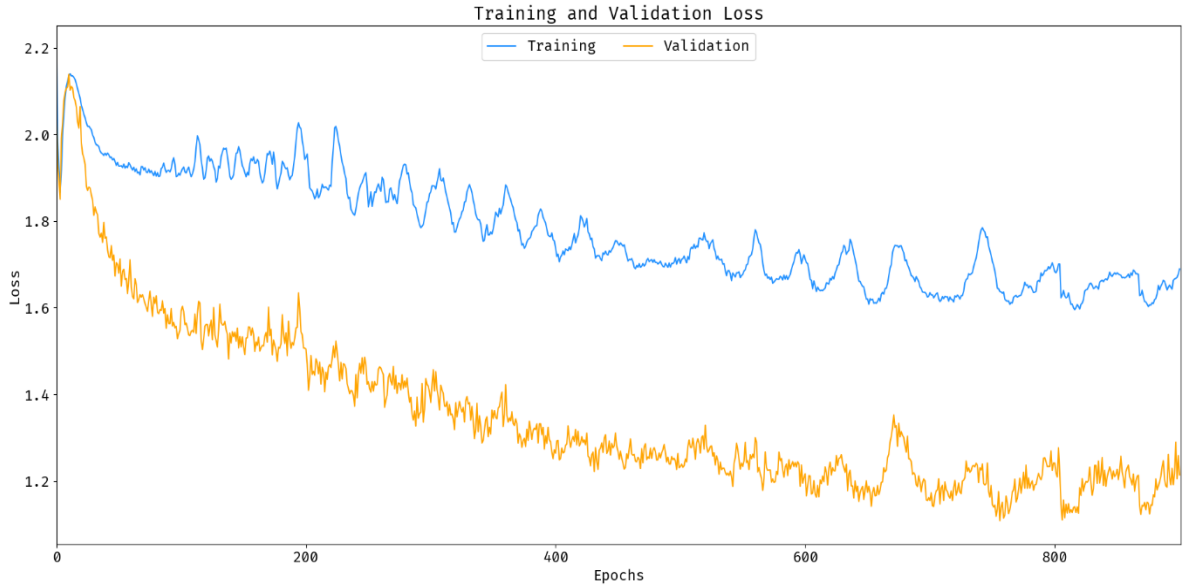

Fig. 7. Loss analysis of IICG model

Table 1 and Fig. 8 provide a detailed comparative outcomes analysis of the proposed technique with existing techniques in terms of different measures. The results have shown that the DVS, AICRL-ResNet50, SS-ENSEMBLE, gLSTM, VD-SAN, Log bilinear, ATT-CNN, and AICRL-VGA16 techniques have obtained minimal outcomes over the other techniques. At the same time, the AoANet, GCN-LSTM+HIP, and M2 (Meshed Memory) transformer techniques have resulted in a moderate outcome. However, the proposed model has gained effective image captioning outcomes with the BLEU-1, BLEU-2, BLEU-3, and BLEU-4 of 83.52, 68.78, 56.32, and 41.53, respectively.

Table 1. Result analysis of IICG model in terms of different measures

Methods	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Log bilinear	65.60	42.40	27.70	17.70
DVS	57.90	38.30	24.50	16.00
AICRL-ResNet50	61.90	45.20	36.80	26.20
AICRL-VGA16	67.20	43.60	33.80	22.50
VD-SAN	65.20	47.10	33.60	23.90
ATT-CNN	66.10	47.20	33.40	23.20
SS-ENSEMBLE	63.90	45.90	31.90	21.70
gLSTM	65.00	46.30	31.70	21.50
AoANet	81.00	65.80	51.40	39.40
GCN-LSTM+HIP	81.60	66.20	51.50	39.30
M^2 Transformer	81.60	66.40	51.80	39.70
IICG	83.52	68.78	56.32	41.53

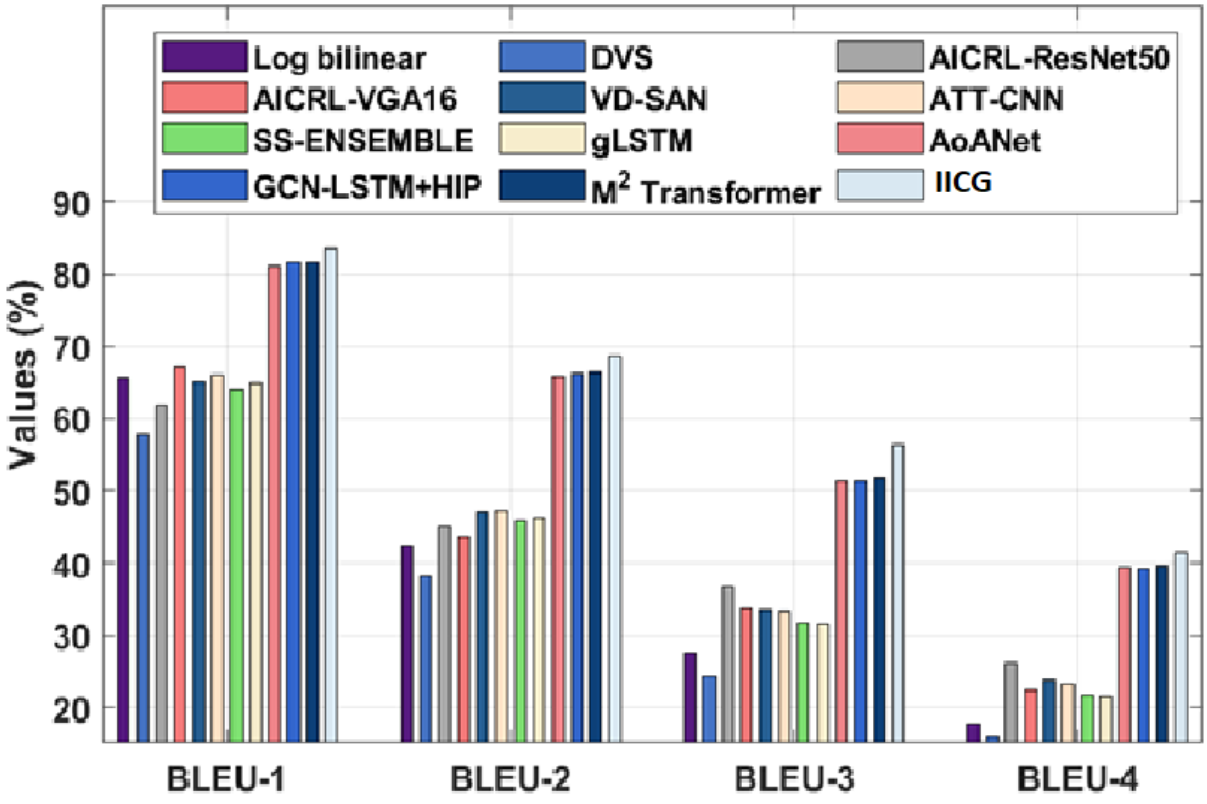

Fig. 8. Result analysis of IICGmodel with varying measures

Table 2 and Fig. 9 investigate the performance of the IICGmodel technique with existing ones in terms of ROUGE-L. The results showcased that the Hard attention and gLSTM models have obtained a lower ROUGE-L of 51.6 and 51.8.

Table 2. ROUGE-L analysis of IICG model with recent approaches

Methods	ROUGE-L
Hard attention	51.60
Adaptive attention	55.00
Semantic attention	53.50
Spatial and channel-wise	53.00
Deliberate attention	58.20
gLSTM	51.80
AoANet	58.90
GCN-LSTM+HIP	59.00
M^2 Transformer	59.20
IICG	61.35

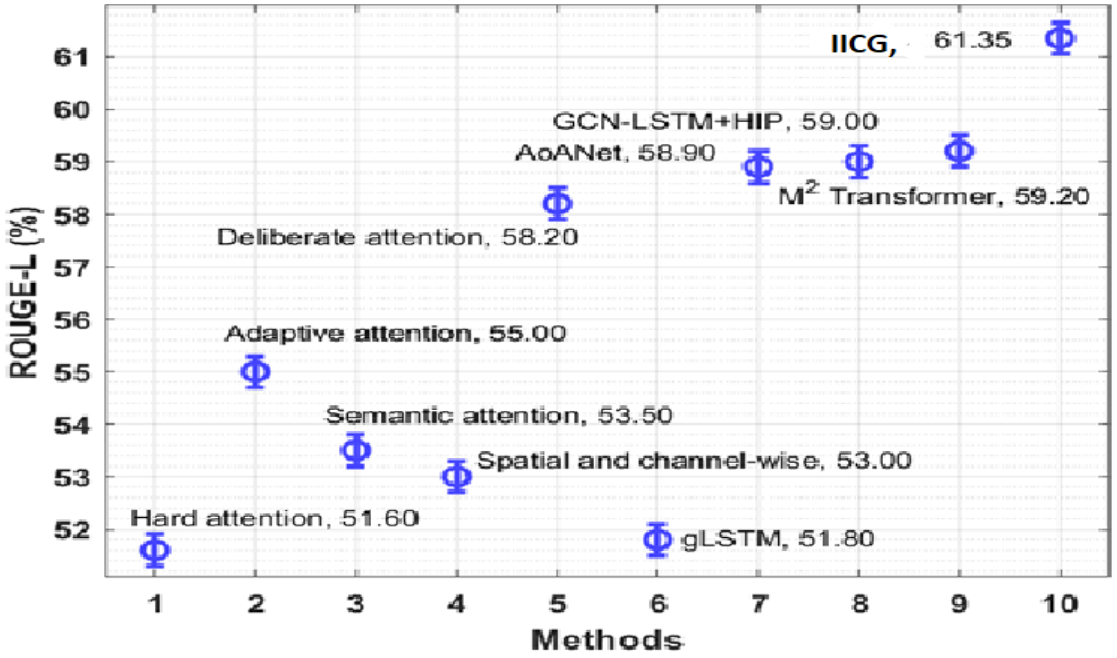

Fig. 9. ROUGE-L analysis of IICGmodel with existing algorithm

Besides, the Spatial and channel-wise, Semantic attention, and Adaptive attention techniques have resulted in a slightly improved ROUGE-L of 53, 53.5, and 55, respectively. Moreover, the Deliberate attention, AoANet, GCN-LSTM+HIP, and M2 Transformer techniques have achieved moderately closer ROUGE-L of 58.2, 58.9, 59, and 59.2, respectively. But, the projected IICGmodel methodology has accomplished an effective outcome with an increased ROUGE-L of 61.35.

Table 3 depicts the comparative analysis of the IICG model with recent approaches with respect to METEOR and CIDEr [24-27]. Fig. 10 explores the performance of the IICG method with recent ones with respect to METEOR. The results exhibited that the log bilinear and AICRL-VGA16 approaches have gained a minimum METEOR of 17.30% and 18.60%. Followed by, the ATT-CNN, VD-SAN, and gLSTM systems have resulted in a slightly higher METEOR of 19.40%, 19.90%, and 20.50% correspondingly. Furthermore, the AICRL-ResNet50, GCN-LSTM+HIP, AoANeT, and M2 Transformer algorithms have obtained moderately closer METEOR of 20.90%, 28.80%, 29.10%, and 29.40% correspondingly. At last, the presented IICGmodel methodology has accomplished effective outcomes with the higher METEOR of 32.88%.

Table 3 Comparative analysis of IICGmodel approach with existing algorithms

Methods	METEOR	CIDEr
Log bilinear	17.30	-
AICRL-ResNet50	20.90	80.30
AICRL-VGA16	18.60	74.30
VD-SAN	19.90	-
ATT-CNN	19.40	-
gLSTM	20.50	54.60
AoANet	29.10	126.90
GCN-LSTM+HIP	28.80	127.90
M^2 Transformer	29.40	129.30
IICG	32.88	133.60

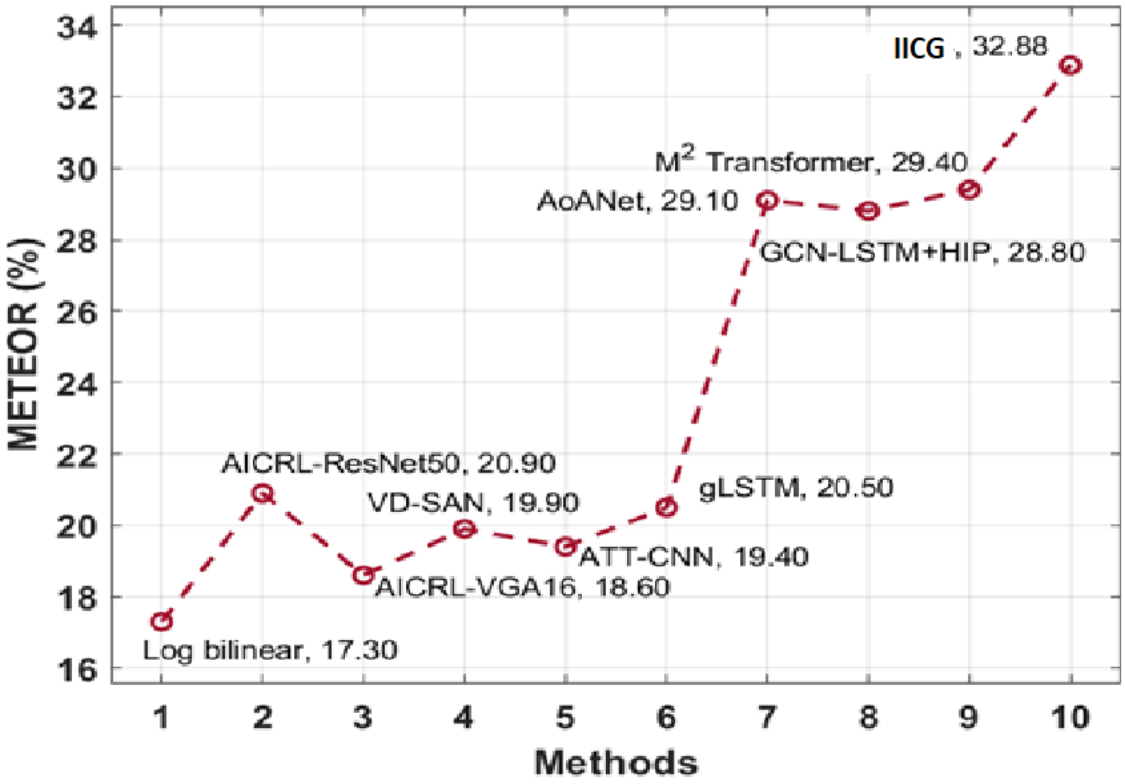

Fig. 10. METEOR analysis of IICGModel with existing algorithm

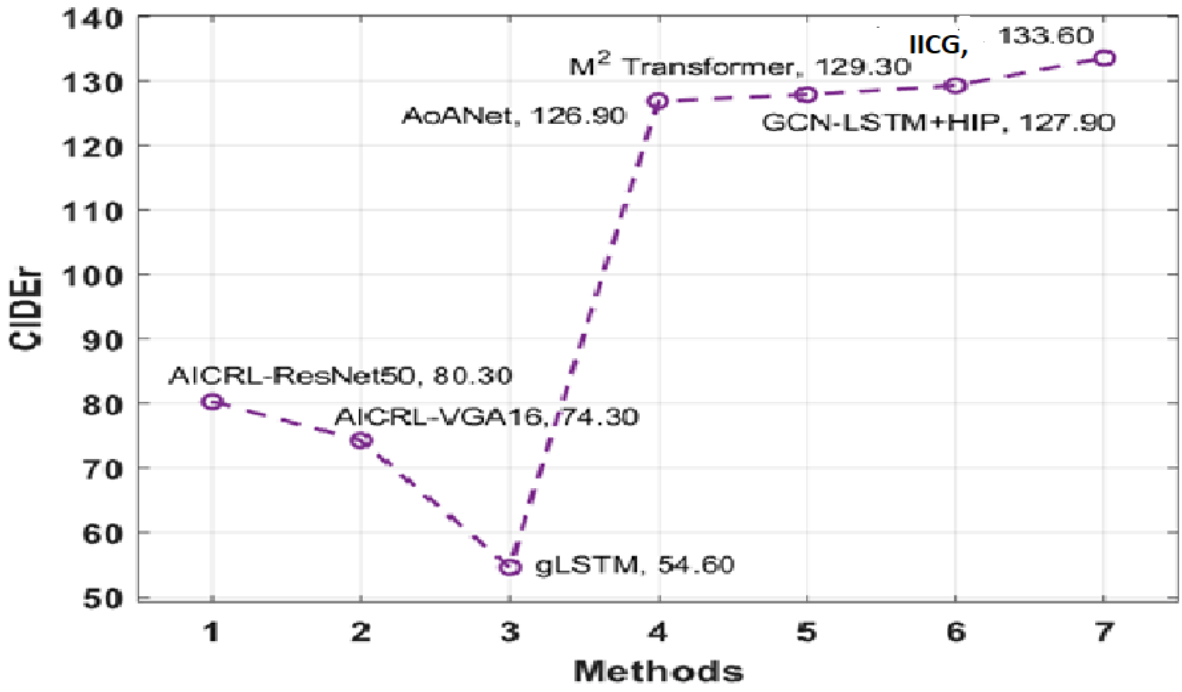

Fig. 11. CIDEr analysis of IICGModel with existing algorithm

Fig. 11 studies the performance of the IICGModel system with present ones in terms of CIDEr. The outcomes outperformed that the gLSTM model has gained the least CIDEr of 54.60. Besides, the AICRL-VGA16 and AICRL-ResNet50 techniques have resulted in a somewhat increased CIDEr of 74.30 and 80.30 correspondingly. Likewise, the AoANet, GCN-LSTM+HIP, and M2 Transformer techniques have achieved moderately closer CIDEr of 126.90, 127.90, and 129.30, respectively. Lastly, the projected IICG system has accomplished effective outcomes with the superior CIDEr of 133.60.

CONCLUSION

In this study, a new IICGmodel has been presented for effective image captioning. The proposed IICGmodel involves preprocessing in several ways, such as removal of punctuation marks, removal of single-letter characters, removal of numerals, and text vectorization. In addition, the features in the image are extracted by the DenseNet121 model, and the image captioning process is performed by the use of Multi-Head Attention Layer, which consists of an encoder as well as a decoder. The simulation analysis of the IICGmodel is carried out against Flickr 8k Dataset. An extensive comparison study is performed with recent approaches, and the obtained results highlighted the better outcomes of the IICGmodel over the existing approaches under several aspects. In the future, the parameter optimization of the DenseNet121 technique can be done by utilizing metaheuristic optimization algorithms.

REFERENCES

[1] P. Anderson, X. He, C. Buehler, et al., Bottom-up and topdown attention for image captioning, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, (2018).

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, Rethinking the inception architecture for computer vision, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016) 2818–2826.

[3] H. Fang, S. Gupta, F. N. Iandola et al., From captions to visual concepts and back, in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015) 1473–1482.

[4] F. Hutter, L. Kotthoff, and J. Vanschoren, Automated machine learning.: methods, systems, challenges, Automated Machine Learning, MIT Press, Cambridge, MA, USA, (2019).

[5] R. Singh, A. Sonawane, and R. Srivastava, Recent evolution of modern datasets for human activity recognition: a deep survey," Multimedia Systems, 26 (2020).

[6] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, HMDB: a large video database for human motion recognition, in Proceedings of the 2011 International Conference on Computer Vision, (2011) 2556–2563.

[7] A. Aker and R. Gaizauskas, Generating image descriptions using relational dependency patterns, in Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 49(9) (2010) 1250–1258.

[8] S. Li, G. Kulkarni, T. L. Berg, and Y. Choi, Composing simple image descriptions using web-scale N-grams, in Proceeding of Fifteenth Conference on Computational Natural Language Learning, 220–228 (2011).

[9] Y. Yang, C. L. Teo, H. Daume, and Y. Aloimonos, Corpusguided sentence generation of natural images, in Proceeding of the Conference on Empirical Methods in Natural Language Processing, (2011) 444–454.

[10] G. Kulkarni, V. Premraj, V. Ordonez, et al., Babytalk: understanding and generating simple image descriptions, IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(12) (2013) 2891–2903.

[11] Chu, Y., Yue, X., Yu, L., Sergei, M. and Wang, Z., Automatic image captioning based on ResNet50 and LSTM with soft

attention. Wireless Communications and Mobile Computing, (2020).

[12] Lu, H., Yang, R., Deng, Z., Zhang, Y., Gao, G. and Lan, R., Chinese image captioning via fuzzy attention-based DenseNet-BiLSTM. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 17(1s) (2021) 1-18.

[13] Zhang, J., Li, K., Wang, Z., Zhao, X., and Wang, Z., Visually enhanced gLSTM for image captioning. Expert Systems with Applications, 184 (2021) 115462.

[14] Zhong, Y., Wang, L., Chen, J., Yu, D., and Li, Y., August. Comprehensive image captioning via scene graph decomposition. In European Conference on Computer Vision (2020) 211-229.

[15] Liu, H., Wang, G., Huang, T., He, P., Skitmore, M., and Luo, X.,. Manifesting construction activity scenes via image captioning. Automation in Construction, 119 (2020) 103334.

[16] Shen, X., Liu, B., Zhou, Y., Zhao, J., and Liu, M., Remote sensing image captioning via Variational Autoencoder and Reinforcement Learning. Knowledge-Based Systems, 203 (2020) 105920.

[17] Del Chiaro, R., Twardowski, B., Bagdanov, A.D. and Van de Weijer, J., Ratt: Recurrent attention to transient tasks for continual image captioning. arXiv preprint arXiv: (2007) 06271.

[18] Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., and Lu, H., Normalized and geometry-aware self-attention network for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020) 10327-10336.

[19] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q., Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (2017) 4700-4708.

[20] Su, S.S., and Kek, S.L., An Improvement of Stochastic Gradient Descent Approach for Mean-Variance Portfolio Optimization Problem. Journal of Mathematics, (2021).

[21] https://www.analyticsvidhya.com/blog/2021/01/implementation-of-attention-mechanism-for-caption-generation-on-transformers-using-tensorflow/

[22] Luo, J. and Ma, L., Image Caption Model based on Multi-head Attention and Encoder-Decoder Framework. In 2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering (ISKE) (2019) 1064-1070. https://www.kaggle.com/adityajn105/flickr8k

[23] Wang, H., Zhang, Y. and Yu, X., An overview of image caption generation methods. Computational intelligence and neuroscience, (2020).

[24] Oluwasammi, A., Aftab, M.U., Qin, Z., Ngo, S.T., Doan, T.V., Nguyen, S.B., Nguyen, S.H., and Nguyen, G.H., 2021. Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning. Complexity, (2021).

[25] Sharma, H. and Jalal, A.S.,. Incorporating external knowledge for image captioning using CNN and LSTM. Modern Physics Letters B, 34(28) (2020) 2050315.

[26] Cornia, M., Stefanini, M., Baraldi, L. and Cucchiara, R., Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 10578-10587.