International Journal of Engineering Trends and Technology
ISSN: 2231 – 5381 /doi:10.14445/22315381/IJETT-V70I3P234
Volume 70 Issue 3, 302-309, March 2022

Original Article

An Attention Mechanism and GRU Based Deep Learning Model for Automatic Image Captioning

Gaurav[1], Pratistha Mathur[2]

[1]*Phd. Scholar, Department of Information Technology, Manipal University Jaipur, Rajasthan , India*
[2]*Professor, Department of Information Technology, Manipal University Jaipur, Rajasthan , India*

[1]gauravsingla31@gmail.com, [2]pratistha.mathur@jaipur.manipal.edu

Abstract *- Image captioning is to generate the image descriptions automatically. In recent years, image captioning has become an active research area and a growing challenge in the field of computer vision and natural language processing. Image captioning using template-based methods and retrieval based approach had some limitations like missing important objects and other attributes. Later on, Encoder-Decoder based methods were presented as research methodologies for image captioning. To extract image information, Convolutional Neural Networks are utilised as encoders. Recurrent Neural Networks are used as decoders to utilise those data and generate content for an image in the encoder-decoder technique. Long short-term memory is the most common recurrent neural network used as a decoder by most researchers. In this paper, a new framework is proposed where Gated Recurrent Unit has been used as a decoder. Along with this, the proposed model has used visual attention for better image features. The proposed framework has been implemented using Flickr8K dataset. The Bilingual Evaluation Understudy score of the proposed framework has been compared with other states of the art frameworks, and it clearly shows that the framework is highly effective and produces state-of-the-art image captions.*

Keywords - *Encoders, Decoders, GRU, Image Captioning.*

I. INTRODUCTION

Image captioning is key research topic of artificial intelligence (AI) that focuses on image interpretation and verbal description. Many platforms like social media, web applications, blogs etc. provide a huge collection of images. The area of computer science that deals with image classification, object detection, video processing is known as computer vision. Natural language processing (NLP) is the technique that deals with understanding human language and makes the computer to behave in the same manner. So here in image captioning, computer vision is used to understand the image and its features. Natural language processing converts those features into text that describes the image. Because of the highly availability of images on social media and other platforms, automatic image captioning has become an active research area. Image captioning can be used in a variety of ways like image indexing, aiding visually impaired people, content-based image retrieval etc. Apart from these real-world applications, there are many fields like biomedicine, commerce, the military, education etc. where image captioning can be used [1].

Many challenges are being faced during image caption generation. The whole process should correctly identify the image, its attributes or properties, such as scene, object, and behaviour. A generated caption should describe the image not only syntactically but also semantically correct. It should be an informative sentence to describe the image appropriately. Image captioning using a template-based method was proposed where a template of fixed size was used to fill up the objects and attributes from the query image. It provided poor performace as it was able to generate only limited sentences with no variance in length of the captions. Image captioning using retrieval approach was proposed in which captions of the same look like images were retrieved, and then one of them was used to select as a caption for the query image. It had some limitations like missing important objects and other attributes. Later on, the deep learning-based encoder decoder method was proposed to overcome previous approaches limitations. To extract image information, CNNs are utilised as encoders. RNNs are used as decoders to utilise those data.

Image features can be extracted using the convolutional layers or fully connected layers of the CNN. Using convolutional layers, local region features can be extracted. Fully connected layers are used to retrieve global features of the image. Most researchers, use fully connected layer only to retrieve the image features and then transfer these features to decoder for the caption generation. But here in the proposed model, image features from convolutional layers have also been retrieved and are transferred to decoder at each step as local image features.

RNN suffers from short term memory issues. RNN doesn't have cell states and uses only hidden states, so RNNs have the memory issues. LSTM is the newer version of RNN and designed to solve short term memory issues and vanishes the gradient problem. LSTM is the

most common RNN used as a decoder by most researchers. But now, Gated Recurrent Unit is a recently developed decoder and has fewer gates than LSTM. GRU uses only two gates instead of three gates as in LSTM. It is less complex than LSTM and is more efficient. It exposes complete memory and hidden layers. LSTM and GRU has been explained in section 3.3. Here in this paper, GRU is used as a decoder which has provided the better results in terms of BLEU (Bilingual Understudy Evaluation) score. BLEU score refers to the match between generated caption and referenced caption. The BLEU score comes in between 0 and 1 and it represents the match between generated caption from the model and referenced caption from the dataset.

The caption generating model takes an image topic pair and generates a caption [2]. Simple encoder and decoder approaches yield good results, but it's challenging to make the most important use of visual information to express picture features or content. The framework proposes an attention based mechanism where visual attention is used to grasp the image better, and GRU has been used a decoder to strengthen the information's integrity. The proposed model has been evaluated on the standard dataset Flickr8K and provides effective results and better BLEU scores. This paper's main contribution or effort is as follows:

- An attention mechanism is applied on CNN to extract local as well as global features.
- GRU is used as a decoder to generate captions.
- The model has been implemented on Flickr8K datasets and the results are highly efficient.

II. RELATED WORK

Image captioning research has been conducted in three dimensions [3]. Image captioning using a template-based method includes a fixed-size template that is to be filled with the objects and attributes from the query image. There are many limitations, like only fixed-size captions can be generated and can ignore the image's semantic representation. Image captioning using retrieval approach is a method in which captions of same look like images are retrieved and then one of them is used to select as a caption for the query image. It has some limitations like missing important objects and other attributes.

Encode decode methods overcome the limitations proposed by earlier approaches. A Convolution Neural Network is a sophisticated learning algorithm that accepts images as input, distributes loads and inclinations, and distinguishes one image from another [4]. Initially, a simple encoder-decoder approach was proposed using CNN and RNN. But RNN suffers from short term memory issues, so LSTM was introduced that worked as a decoder and removed short term memory issues.

There is a method of avoiding words with identical appearances but different meanings by employing negative sampling instances and really difficult negative examples [5]. An author offers a sound continuity module based on gated recurrent units [6]. In one research, the author suggests strategies for producing different captions to develop specific and thorough captions [7]. Multi-task learning is emphasised in a model, which helps to improve model generality and performance [8]. For picture captioning, The Deep Hierarchical Encoder Decoder Network was introduced, which can efficiently integrate vision and language semantics at a high level in the construction of captions by utilising deep networks' representation capability [9]. One approach uses adaptive and dense net attention mechanisms [10]. A model presents a domain-specific picture caption generator that combines object and attribute information with attention mechanisms to create captions using a semantic ontology to deliver natural language descriptions for a specified domain [11]. A twofold attention model combining an attention model at the sentence level and an attention model at the word level has been developed to construct more accurate captions [12]. One method relies on a framework that employs an attention balance mechanism as well as a syntax optimisation module [13]. In the next paper, the author offers a model that tightly combine attribute recognition with image captioning and urge successful attribute usage by predicting acceptable attributes at each step [14]. The RNN's topological inner structure is approximated for image captioning [15]. The process of sentence generation is viewed as a problem in which the proper sentence generating probability is maximised based on the picture information provided [16]. A Hierarchical Attention Fusion model, which combines Resnet multi-level feature maps with hierarchical attention, is presented as a baseline for image captioning based on RL [17]. To avoid the vanishing gradient problem, LSTM is an RNN variation [18]. For picture captioning, the Integrated Dual Generative Adversarial Network (IDGAN) combines retrieval and generation approaches [19]. One idea for the remote sensing picture captioning problem is to use a Reinforcement Learning with a Variational Autoencoder Learning Model with Two Stages and Multiple Tasks [20]. One model is proposed with multiple encoders and decoders using multihead attention layer transformer model [21]. In one of the paper, objects are detected using thermal images that may be useful in image captioning as object detection is done at encoder part in image captioning process [22].

This framework proposes an attention based mechanism where visual attention is used to grasp the image better, and GRU has been used a decoder to strengthen the information's integrity. The proposed model has been put to the test on the standard dataset Flickr8K and provides effective results and better BLEU scores.

III. THE PROPOSED MODEL

Here in this section, the proposed model has been presented step by step. The proposed model consists of four main components:

- Inception ResNetV_2 has been used as an encoder.
- Mechanism of visual attention is achieved by extracting local region features from

- convolutional layers of the encoder.
- Global features of the image are retrieved using last fully connected layer of the encoder.
- GRU has been used as a decoder which is latest version of RNN.

A. Problem formulation

The image (M) and its referenced statement (C) will be given, and during the model training phase, the maximum likelihood is utilised to increase the likelihood of providing an image caption as shown in Eq. 1.

$$W^* = argmax\left(\sum_{M,C} \log(p(C|M;W)\right) \qquad \text{Eq. 1}$$

where W stands for the parameter that needs to be learned and is trained by the image caption generating model.

During the training, the weight parameters will be learned in such a manner that the caption C can be generated for the given image M. The pictorial representation of the model is shown in Fig. 1.

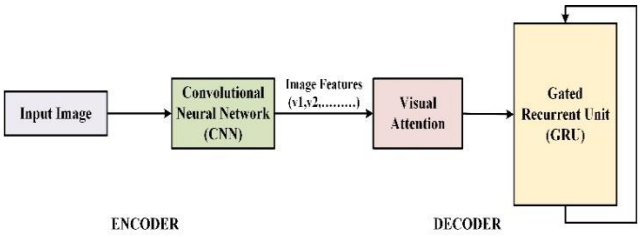

Fig. 1 Double Awareness Based Model for Image Caption Generation

B. Image caption generation using attention based mechanism

During the model training phase, the encoder receives images as inputs and the decoder is trained in such a manner that referenced captions can be generated during testing phase. The image's overall features are extracted using last fully connected layer of Convolutional Neural Network and local features are extracted using last convolutional layer. The overall features of the image are transferred to the decoder only at initial time step as initial hidden state of the network. After this, local image features are applied on decoder for better understanding of the image at next time steps. Image feature vectors are mapped with the input part of GRU. During caption generation, image feature vector, input word that has been generated at previous time step and generated hidden state at previous time step are applied as input to the decoder. The decoder generates next word and next hidden state and these outputs will be used as input again at next time step. This process is repeated until the desired caption is generated. The dataset that has been used for the implementation is Flickr8K. Flickr8K is the dataset that contains nearly 8000 images, and there are five captions per image. Out of 8000 images, 6000 images have been used as training images, 1000 images have been used as testing images and the remaining 1000 images have been used as validation images. There are 40,000 captions for these 8000 images. As global features alone are not enough to extract regional image features, local features have also been extracted using Convolutional Neural Network. This paper uses the predefined training model Residual Network (ResNet) as a Convolutional Neural Network. As a result, the proposed model is based on an attention based mechanism, in which visual attention mechanisms are used to generate image captions. Apart from this, GRU has been used as a decoder that has provided highly efficient results.

For the given image M and descriptive sentence S={S_0, S_1, S_2,……, S_N}, the Gated Recurrent Model is trained, and the training process takes place as follows:

$$f_g = CNN(M) \qquad \text{Eq. 2.}$$

$$x_t = Lookup(w_e, s_t), t\in\{0,1,2,\ldots\ldots,N\} \qquad \text{Eq. 3.}$$

$$h_t = GRU(x_t, v_t) \qquad \text{Eq. 4.}$$

$$s_{t+1} \sim r_{t+1} = softmax(h_{t+1}) \qquad \text{Eq. 5.}$$

Using CNN, global features f_g are extracted at the initial step as shown in Eq. 2. Initially, these global features are used to activate the GRU model. w_e is the word embedding matrix for all words and x_t i.e. word vector can be retrieved by looking up the word matrix for each word s_t as shown in Eq. 3. For every iteration, the current word vector x_t and visual attention vector v_t are combined and passed as input to Gated Recurrent Unit i.e., GRU. It will generate the current hidden state as shown in Eq. 4. Then, SoftMax function is used to generate selection probability vector r_t from this hidden state. The word with maximum probability is selected and passed as input to the next step as shown in Eq. 5. Here v_t is the visual attention vector that is extracted using last convolutional layer of CNN. So, at every time step these three vectors x_t, v_t are updated and passed as input to GRU to generate the next word for the caption.

C. LSTM vs GRU

Although RNN-based algorithms can generate reasonable phrases, they can cause vanishing gradients during training [23]. LSTM can resolve this issue and its core working is based on its cell states and various gates. The gates carry the relevant information that may be used later for sentence generation or some other tasks. Important information received at earlier time steps may be used later also as it keeps such information in memory. The gates are used to retain the information and forget the information required later on. It uses the sigmoid activation function to use values as 0 or 1. Here 0 means the information is not required for future use and, 1 means the information is to be retained and will be used for future tasks.

As mentioned earlier, three gates are used by LSTM. The forget gate is used to determine whether the information is to be retained or forgotten. Input vector and information from previous hidden states are applied on sigmoid activation function and, it provides results in the form of 0 or 1. Based on the resulted value, the information is either kept or forgotten. Input gate is used to update the information. Input vector and previously hidden state information are passed to the sigmoid activation function and the tanh activation function. The output from activation function (sigmoid) is used to decide which information is required to keep from output of tanh activation function. The purpose of this gate is to update the information. LSTM architecture is shown in Fig. 2.

Once the forget gate and input gate is ready to provide their cell information, new cell states can be easily determined. Cell state is pointwise multiplied by forget vector. Then the output of this pointwise multiplication is added with the output of the input gate, and new cell states are received. This way, updating of cell states is performed. Output gate is used to generate the next hidden state. Like other gates, the current input vector and previous hidden state are applied to the activation function (sigmoid). New cell state is applied on activation function (tanh). The outputs of both activation functions are compounded, resulting in a new hidden state.

As per LSTM architecture, it can be clearly seen that input vector and previous hidden states are concatenated. Then this concatenation is passed to the forget layer that releases the data that is no longer used for future reference as shown in Eq. 6. Input gate is used for update purpose. It performs pointwise multiplication of outputs received from sigmoid and tanh of the concatenation of current input vector and previous hidden state. These calculations have been shown in Eq. 7 and Eq. 8. Output from forget layer and input layer are combined to calculate a new state as shown in Eq. 9. Using previous hidden state and current input vector and sigmoid activation function and output state is obtained as shown in Eq. 10. Pointwise multiplication of output and cell state provides the new hidden state as shown in Eq. 11. The generated output is used as a previous hidden state in the next step. A new input vector is applied as inputs, and the same process is repeated to generate the output or as per the task's requirement to be solved.

.

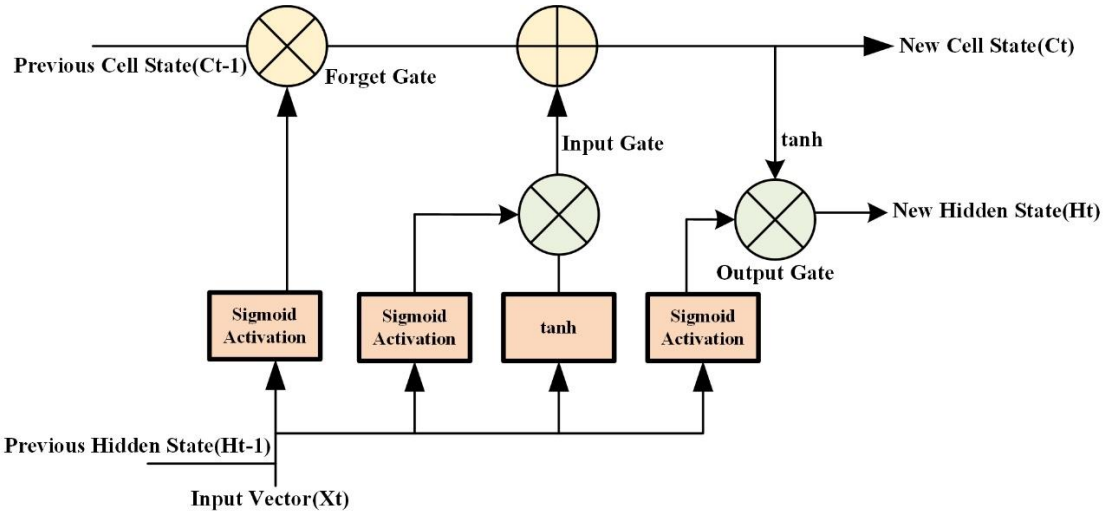

Fig. 2 LSTM Architecture

The following are the equations of LSTM for different gates like input gate, output gate and forget gate:

$$F_t = \sigma\left(W_{forget}[h_{t-1}, x_t] + bias\right) \quad \text{Eq. 6}$$

$$I_t = \left(W_{input}[h_{t-1}, x_t] + bias\right) \quad \text{Eq. 7}$$

$$C_a = tanh\left(W_{input}[h_{t-1}, x_t, v_t] + bias\right) \quad \text{Eq. 8}$$

$$C_t = f_t * C_{t-1} + I_t * C_a \quad \text{Eq. 9}$$

$$O_t = \sigma\left(W_{output}[h_{t-1}, x_t] + bias\right) \quad \text{Eq. 10}$$

$$H_t = O_t * \tanh(C_t) \quad \text{Eq. 11}$$

The proposed framework is tested on GRU methods and compared with other LSTM based existing methods. The model provided better results in the case of GRU as it performs the same task using two gates only and consumes full memory, so it performs better caption generation compared to the LSTM model.

GRU works similar to LSTM but uses only two gates instead of three gates. It uses a reset gate and update gate. The reset gate decides how much information from the past should be erased. It sounds similar to update gate, but both gates are technically different and used by GRU for tasks like caption generation. The proposed model has performed better on GRU as compared to the LSTM. In the next section, the experimental results are shown performed on

the Flickr8k dataset using both LSTM and GRU approaches for image caption generation. It is clearly seen that GRU has performed better on Flickr8K dataset for image captioning task. Fig. 3 depicts the GRU architecture. Here, update gate and reset gates are used for keeping the important information and forget the information that is not required. The concept of keeping important words and forgetting the words which are no more required is very important to generate new words and this work is done easily by using two gates only, so GRU provided better results in the proposed model. In the next section, the results have been compared using standard datasets for image captioning and can be clearly seen that the GRU provided the best results.

As GRU is newer version of RNN than LSTM, it has some key differences from LSTM.:

- Unlike the LSTM, which has three gates, GRU only has two gates.
- GRU requires less training parameters and is more efficient computation wise. As a result, it is less difficult to train than LSTM [24].
- GRU exposes complete memory and hidden layers.
- GRU was invented in 2014 and LSTM was invented in 1995-97.

The following are the working equations for GRU:

$$Z(t) = \sigma\left(W_Z . X_t + U_z . h_{t-1} + bias_{update}\right) \quad \text{Eq. 12}$$

$$R(t) = \sigma(W_R . X_t + U_r . h_{t-1} + bias_{reset}) \quad \text{Eq. 13}$$

$$h' = tanh\left(W_h . X_t + R(t) * U_h . h_{t-1} + bias_{update}\right) \quad \text{Eq. 14}$$

$$h = Z(t) * h_{t-1} + \left(1 - Z(t)\right) * h' \quad \text{Eq. 15}$$

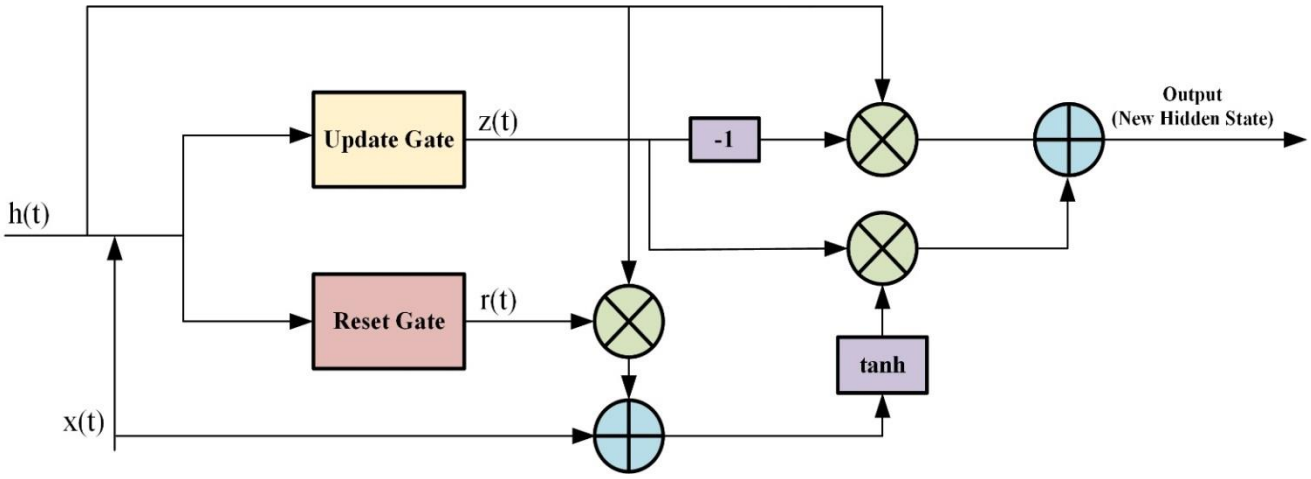

Fig. 3 GRU Architecture

GRU vanishes the gradient problem of RNN and is highly useful for applications wherein sentence generation is involved. At time step t-1, weight vectors of the update gate are pointwise multiplied by the input vector, and hidden state is pointwise multiplied by the Update gate vector. These two vectors are added along with the bias vector. The sigmoid function is applied to the overall result to generate a new update vector for time step t as mentioned in Eq. 12. Similarly, the input vector is pointwise multiplied by the weight vector of the reset gate and the hidden state is multiplied by reset gate vector. Then these two vectors are added along with bias vector. The sigmoid function is applied on the overall result to generate a new reset gate vector for time step t as shown in Eq. 13. The gate vector is reset and updated for time step t.

Hidden state weight vector is pointwise multiplied by the input vector. The hidden state vector is pointwise multiplied by update vector and normal multiplication by newly generated reset vector. The result is added into bias vector of update. It generates partially new hidden state or output state as shown in Eq. 14. This partially generated new hidden state is multiplied by the negation of the update vector. Previous hidden state is multiplied by update vector. These two vectors are added to generate the output of GRU at time step t as shown in Eq. 15. The output at time step t is also the new hidden state.

IV. EXPERIMENTAL RESULTS

The most famous datasets for image captioning are MSCOCO, Flickr30K and Flickr8K. Flickr30k has a total of 31, 783 images. One of the most common datasets for image captioning is MS COCO. There are a total of 123,287 photos in this collection [25]. MSCOCO dataset contains more than 3,00,000 images based on various objects and their properties. Flickr30K dataset contains 30,000 images and there are 5 captions per image. So overall Flickr30K contains 1,50,000 captions for 30K images. Flickr8K dataset contains 8,000 images and there

are 5 captions per image. So overall, there are 30000 captions for all images in Flickr8K dataset. Apart from these datasets, visual genome dataset, Instagram dataset, Stock3M dataset, Flickr style 10k dataset etc. These datasets are highly efficient for image captioning, including instances on various categories. From the datasets, images are used for the training, testing and validation. In this paper, the dataset used for the model is Flickr8K. The summary of Flickr8k dataset has been shown in Table 1. From 8,000 images, 6,000 images have been used for training purpose. 1,000 images have been used for testing and remaining 1,000 images for validation purpose. However, pre-processing of the captions has been done to ignore the words that may no longer help generate better sentences for the query images.

Various evaluation metrics are used for image captioning purpose. In this paper, the model has been evaluated on the BLEU score, METEOR score, ROUGE score, CIDEr score. BLEU score refers to match between generated caption and referenced caption. BLEU score comes in between 0 and 1. The score 0 means there is no matching between generated and referenced caption and score 1 means all words are matched between generated and referenced caption. There are four different scores of BLEU metrics such as BLEU-1, BLEU-2, BLEU-3 and BLEU-4. BLEU-1 score shows the matching of one word or gram in between generated and referenced caption. BLEU-2 score shows the matching of two words or grams in between generated and referenced caption. BLEU-3 score shows the matching of three words or grams in between generated and referenced caption. BLEU-4 score shows the matching of four words or grams in between generated and referenced caption. All scores have been evaluated for the proposed model. METEOR evaluation metric is mainly used to focus on synonyms of the words. Only BLEU score is not sufficient to specify the quality of generated caption. Apart from these two metrics, ROUGE and CIDEr have also been evaluated and presented in resultant matrix. Table 2 shows the model's experimental results on Flickr8K Dataset.

Table 1. Flickr8k dataset

	Images in Dataset	**Training size**	**Testing Size**	**Development Size**
Flickr8K Dataset	8,092 Images	6,000 Images	1,000 Images	1,000 Images
	Captions per image	**Captions in Dataset**	**Unique words**	**Max length of any caption**
	5	40,000	8763	40 Words

Table 2. Experimental result using various encoders as decoders

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE	CIDEr
NIC(InceptionV3) using LSTM [26]	0.725	0.596	0.488	0.402	0.346	0.597	1.107
NIC(InceptionV4) using LSTM [26]	0.721	0.590	0.483	0.398	0.344	0.592	1.131
Proposed Model (ResNet and GRU along with attention mechanism)	0.742	0.618	0.508	0.419	0.372	0.605	1.109
An Attention Model (InceptionV3 and GRU)	0.731	0.601	0.496	0.415	0.361	0.601	1.098

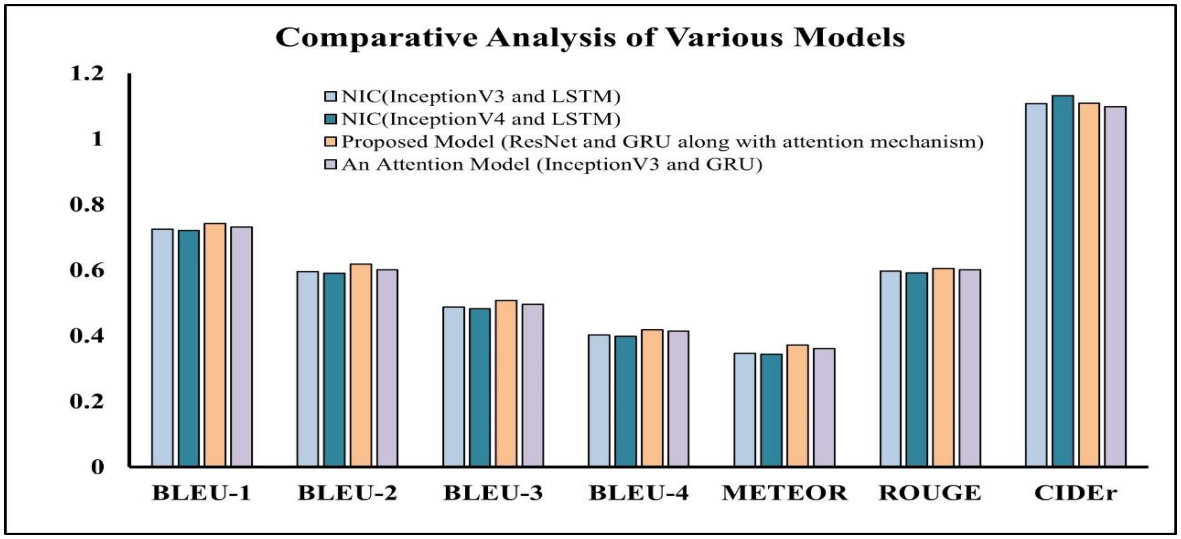

Fig. 4 Comparative Analysis of models using various encoders and decoders

From the comparative analysis as shown in Fig. 4, it can be found that GRU has provided better results than LSTM for the image captioning model. The proposed model has been tested on ResNet and Inception V3 as encoders and LSTM as well as GRU as decoders. The proposed model attains attention mechanism, and better results are achieved when the combination of ResNet and GRU has been used. Using attention mechanism, local features of the image are used at every time step of language model. Global features of the image are transferred only once but local features are used at every time step during word generation by language model. In Table 2, NIC(InceptionV3) and NIC(InceptionV4) models use LSTM as decoder, and improved results can be seen when GRU is used as a decoder. BLEU-1 score shows the generated caption matches the referenced caption from the dataset. High accuracy means almost likely to reference caption. The model has focused on the visual attention to generate the caption for the given query image.

In this paper, the experiments have been conducted using Flickr8K dataset. One model is attention-based model that uses InceptionV3 as encoder and GRU as the decoder. In this model, the BLEU scores are 0.731, 0.601, 0.496, 0.415 respectively. Another model is attention based using ResNet as encoder and GRU as a decoder where BLEU scores can be seen as 0.742, 0.618, 0.508, 0.419 respectively . Among all these models, it can be found that the proposed model using ResNet as encoder and GRU as decoder has provided the best result. In Fig. 5, some image caption examples are generated using the proposed model on the Flickr8K dataset. Generated captions are highly relevant to the images, and sentences are grammatically and semantically correct.

Fig. 5 Image Caption Generation using Proposed Model

V. CONCLUSION

In the paper, image captioning has been implemented using the encoder and decoder approach along with attention based mechanism. Visual attention has been used to obtain essential features of the image. Extracting only global features may ignore some important attributes, so here visual attention has been used that provides local features of the image and are helpful in identifying local region based objects and their properties. ResNet and GRU has been used as an encoder and decoder, respectively. GRU works similar to LSTM but uses only two gates instead of three gates. It uses reset gate and update gate. It

is less complex and exposes complete memory and hidden layers. The framework has been tested on Flickr8K dataset. The proposed framework using GRU as a decoder has provided a better BLEU score than other frameworks that used LSTM as a decoder. BLEU-1, BLEU-2, BLEU-3 and BLEU-4 scores are better in the framework. As a future work, the more focus will be on textual attention mechanism so that double awareness mechanism can be created along with optimizing important parameters.

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. 51(6) (2019). doi: 10.1145/3295748.

[2] N. Yu, X. Hu, B. Song, J. Yang, and J. Zhang, Topic-Oriented Image Captioning Based on Order-Embedding, IEEE Trans. Image Process. 28(6) (2019) 2743–2754. doi: 10.1109/TIP.2018.2889922.

[3] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, Boosting Image aptioning with Attributes.

[4] K. Loganathan, R. Sarath Kumar, V. Nagaraj, and T. J. John, CNN & LSTM Using Python for Automatic Image Captioning, Mater. Today Proc. (2020). doi: 10.1016/j.matpr.2020.10.624.

[5] W. Cai and Q. Liu, Image Captioning with Semantic-Enhanced Features and Extremely Hard Negative Examples, Neurocomputing. 413 (2020) 31–40. doi: 10.1016/j.neucom.2020.06.112.

[6] X. Lu, B. Wang, and X. Zheng, Sound Active Attention Framework for Remote Sensing Image Captioning, IEEE Trans. Geosci. Remote Sens. 58(3) (2020) 1985–2000. doi: 10.1109/TGRS.2019.2951636.

[7] Y. Jing, X. Zhiwei, and G. Guanglai, Context-Driven Image Caption with Global Semantic Relations of the Named Entities, IEEE Access. 8 (2020) 143584–143594. doi: 10.1109/ACCESS.2020.3013321.

[8] C. Wang, H. Yang, and C. Meinel, Image Captioning with Deep Bidirectional LSTMs and Multi-Task Learning, ACM Trans. Multimed. Comput. Commun. Appl. 14(2s) (2018). doi: 10.1145/3115432.

[9] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, Deep Hierarchical Encoder-Decoder Network for Image Captioning, IEEE Trans. Multimed. 21(11) (2019) 2942–2956. doi: 10.1109/TMM.2019.2915033.

[10] Z. Deng, Z. Jiang, R. Lan, W. Huang, and X. Luo, Image Captioning Using Densenet Network and Adaptive Attention, Signal Process. Image Commun. 85 (2020) 115836. doi: 10.1016/j.image.2020.115836.

[11] S. H. Han and H. J. Choi, Domain-Specific Image Caption Generator with Semantic Ontology, Proc. - 2020 IEEE Int. Conf. Big Data Smart Comput. Bigcomp. (2020) 526–530. doi: 10.1109/BigComp48618.2020.00-12.

[12] H. Wei, Z. Li, C. Zhang, and H. Ma, The Synergy of Double Attention: Combine Sentence-Level and Word-Level Attention for Image Captioning, Comput. Vis. Image Underst. 201 (2019) 103068. doi: 10.1016/j.cviu.2020.103068.

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

[14] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, Image Captioning with End-to-End Attribute Detection and Subsequent Attributes Prediction, IEEE Trans. Image Process. 29 (2020) 4013–4026. doi: 10.1109/TIP.2020.2969330.

[15] H. Wang, H. Wang, and K. Xu, Evolutionary Recurrent Neural Network for Image Captioning, Neurocomputing. 401 (2020) 249–256. doi: 10.1016/j.neucom.2020.03.087.

[16] X. Lu, B. Wang, X. Zheng, and X. Li, Sensing Image Caption Generation, IEEE Trans. Geosci. Remote Sens. 56(4) (2017) 1–13.

[17] C. Wu, S. Yuan, H. Cao, Y. Wei, and L. Wang, Hierarchical Attention-Based Fusion for Image Caption with Multi-Grained Rewards, IEEE Access. 8 (2020) 57943–57951. doi: 10.1109/ACCESS.2020.2981513.

[18] L. Gao, X. Li, J. Song, and H. T. Shen, Hierarchical LSTMs with Adaptive Attention for Visual Captioning, IEEE Trans. Pattern Anal. Mach. Intell. 42(5) (2020) 1112–1131. doi: 10.1109/TPAMI.2019.2894139.

[19] J. Liu et al., Interactive Dual Generative Adversarial Networks for Image Captioning, AAAI 2020 - 34th AAAI Conf. Artif. Intell. (2020) 11588–11595. Doi: 10.1609/aaai.v34i07.6826.

[20] X. Shen, B. Liu, Y. Zhou, J. Zhao, and M. Liu, Remote Sensing Image Captioning Via Variational Autoencoder and Reinforcement Learning, Knowledge-Based Syst. 203 (2020) 05920. doi: 10.1016/j.knosys.2020.105920.

[21] J. Jansi Rani and B. Kirubagari, An Intelligent Image Captioning Generator using Multi-Head Attention Transformer, Int. J. Eng. Trends Technol. 69(12) (2021) 267–279. doi: 10.14445/22315381/IJETT-V69I12P232.

[22] V. Teju and D. Bhavana, An Efficient Object Tracking in Thermal Imaging Using Optimal Kalman Filter, Int. J. Eng. Trends Technol. 69(12) (2021) 197–202. doi: 10.14445/22315381/IJETT-V69I12P223.

[23] B. Wang, X. Zheng, B. Qu, X. Lu, and S. Member, Remote Sensing Image Captioning, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sensing. 13 (2020) 256–270.

[24] B. C. Mateus, M. Mendes, J. T. Farinha, R. Assis, and A. M. Cardoso, Comparing LSTM and GRU Models to Predict the Condition of a Pulp Paper Press, Energies. 14(21) (2021) 1–21. doi: 10.3390/En14216958.

[25] S. Kalra and A. Leekha, Survey of Convolutional Neural Networks for Image Captioning, J. Inf. Optim. Sci. 41(1) (2020) 239–260. doi: 10.1080/02522667.2020.1715602.

[26] M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, Image Caption Generation with Dual Attention Mechanism, Inf. Process. Manag. 57(2) (2020) 102178. doi: 10.1016/j.ipm.2019.102178.