A Comparative Study of Pre-trained CNNs and GRU-Based Attention for Image Caption Generation

Rashid Khan*
University of Science and Technology of China, Hefei
College of Big Data and Internet SZTU Shenzhen, China
Email: rashidkhan@mail.ustc.edu.cn

Bingding Huang*
College of Big Data and Internet
SZTU Shenzhen, China
Email: huangbingding@sztu.edu.cn

Haseeb Hassan
College of Health Sciences
and Environmental Engineering
SZTU Shenzhen, China
Email: Haseeb@sztu.edu.cn

Asim Zaman
College of Health Sciences
and Environmental Engineering
SZTU Shenzhen, China
Email: zamanasim2021@email.szu.edu.cn

Zhongfu Ye
Department of Electronic Engineering
and Information Science
University of Science and Technology of China, hefei
Email: yezf@ustc.edu.cn

***Abstract*—Image captioning is a challenging task involving generating a textual description for an image using good computer vision and natural language processing techniques. This paper proposes a deep neural framework for image caption generation using a GRU-based attention mechanism. Our approach employs multiple pre-trained convolutional neural networks as the encoder to extract features from the image and a GRU-based language model as the decoder to generate descriptive sentences. To improve performance, we integrate the Bahdanau attention model with the GRU decoder to enable learning to focus on specific image parts. We evaluate our approach using the MSCOCO and Flickr30k datasets and show that it achieves competitive scores compared to state-of-the-art methods. Our proposed framework can bridge the gap between computer vision and natural language and can be extended to specific domains.**

***Index Terms*—Image captioning, Attention mechanism, Inception V3, Convolutional Neural Network, GRU.**

I. Introduction

To create descriptive sentences for images, image caption generation is a critical challenge in computer vision (CV) and natural language processing (NLP). Recent advancements in caption generation have primarily revolved around the encoder-decoder framework. This architecture employs a neural network as the encoder to extract features from the source image, while a language model serves as the decoder to generate the target sentence. Though this approach has displayed promising outcomes, attention mechanisms are essential for assessing the significance of the encoder's hidden states. However, conventional sequential attention mechanisms lack comprehensive modeling capabilities, necessitating a review network to create concise and abstract annotation vectors.

On the other hand, NLP has spawned a slew of applications ranging from primary text classification to fully automated natural language chatbots. It has been a critical and fundamental challenge in the Deep Learning domain. Image captioning (IC) has a wide range of applications, including (i) transcribing scenes for people who are blind [1], [2], (ii) classifying videos and photographs based on various situations [3], (iii) image-based search engines for better results, [4] (iv) visual question answering [5], and (v) context comprehension [6].

To generate the corresponding sentence for a given image, the latest research on caption generation, such as image captioning [7], relies on an encoder-decoder framework. Different neural network architectures are employed as the encoder due to the other behavior and features of the source, like convolutional neural networks (CNNs) for images and recurrent neural networks (RNNs) for sequential data, including source code and natural language. The attention technique evaluates the significance of the encoder's hidden states based on all previously generated words in theOnme step. The attache attention mechanism works for sequential hand, works in a sequential manner and lacks global modeling capabilities. A review network [8] was proposed to overcome this flaw, with review steps located here between the encoder and the decoder. Conventional techniques for generating image captions often involve assembling, hindering the smooth information exchange change of information between visual and language realms. Recurrent Neural Networks (RNNs) have exhibited prowess in crafting sequences, as evidenced in natural language processing. Meanwhile, Convolutional Neural Networks (CNNs) excel at extracting features from images. Due to the distinct attributes of each domain, fusing these two in image captioning has posed considerable challenges. To overcome this, our research proposes an innovative integrated approach that harmonizes CNNs and Gated Recurrent Units (GRUs) by utilizing attention networks. The underlying principle is rooted in the understanding that a comprehensive model adept at seamlessly melding visual feature extraction with contextual text generation can substantially enhance the quality of resultant captions. The following are the main contributions of our paper:

*Corresponding author: Rashid Khan (rashidkhan@sztu.edu.cn) and, Bingding Huang (huangbingding@sztu.edu.cn)

979-8-3503-2635-2/23/$31.00 ©2023 IEEE

- For image caption generation, we examined the encoder-decoder framework. The ENCODER would use a pre-trained Convolutional Neural Network (CNN) to encode the image, and the DECODER would use a Recurrent Neural Network (RNN) to create each caption word iteratively.
- The model's performance was compared to four pre-trained CNNs: InceptionV3, DenseNet169, ResNet101, and VGG16. We employ the GRU with soft attention as the decoder, effectively focusing the attention over a specific part of an image to predict the next sentences.
- In our image captioning approach, we apply an attention mechanism that can focus on the essential elements of the image and define fine-grained captions. Finally, we utilize the open-source MSCOCO [9] and Filker30k datasets to quantitatively validate the research paper's utility in image caption generation.

II. Related Work

In related work, we enhance relevant information on a prior study on image caption generation and attention. Several approaches for generating image descriptions have recently been presented. Automatic image captioning generation has emerged as a promising research area in recent years because of advances in deep neural network models for CV and NLP. In general, there are three types of image captioning modeling techniques: neural-based approaches [10]–[12], attention-based strategies [13]–[16], and RL-based methods framework [17], [18]. Attention-based approaches have recently gained popularity and are more successful than neural-based methods. When guessing each word in the caption, attention-based techniques focus on specific locations in the image. Deep neural networks (DNNs) were initially proposed for caption generation [19]. Integrating context-rich information using RGB-D descriptors improves visual comprehension by integrating color and depth data [20]. It also considers efficiency by utilizing filter pruning techniques within CNNs to reduce computational complexity, providing practical real-time image captioning [21]. Furthermore, advanced clustering methods for semantic grouping are being investigated, to produce more coherent and informative captions via improved K-means cluster quality, contributing to a comprehensive approach that embraces cutting-edge advancements in computer vision and machine learning [22]. Some are comparable instances of similar work [23] that utilize CNN and RNN to generate descriptions. Visual attention is an efficient image caption generation approach [24], [25]. When developing the target language, these attention-based captioning models may learn where to focus on the image. They may understand the distribution of spatial attention during the last convolutional layer of the CNN [26], or they may learn the distribution of semantic attention from visual characteristics known from social media images [27]. Whereas these methodologies demonstrate the efficiency of the attention mechanism, they do not investigate the contextual information in the encoding sequence. Our attention layer is distinct in that it is structured in sequential order, with each hidden state of an encoding stage contributing to the formation of decoding words. An Attention mechanism can be used to improve the contextual aspect of natural language sequences. The use of attention to describing image content is consistent with human understanding [24]. The evaluation matrix and the accuracy of attention to imaging correlate significantly. Even still, the measure to which attention accuracy is congruent with human perceptions needs to be increased [28]. The attention area captioning model comprises image regions, word captions, and the NLP natural language framework (RNN). The MS COCO dataset is typically used to test the trained system [29]. In artificial intelligence, developing a caption that accurately represents an image is essential [30]. One of the procedures in image captioning is extracting coherent characteristics of an image using an image-based framework. The extracted features of the image-based model are used to describe an image in NL. We suggested creating an image captioning model that employs a GRU with a soft attention decoder to predict future sentences by selectively focusing attention on a specific part of an image. We used cutting-edge architecture to evaluate the model's performance to that of four pre-trained CNNs: Inception V3, DenseNet169, ResNet101, and VGG16. The Attention layer is used to make the image's caption more sensible.

III. GRU-Based Attention Network

Extracting visual information and expressing it in a grammatically correct natural language sentence are the two main components of automatically generating natural language sentences that describe an image. Figure III shows a simple Encoder-Decoder deep learning-based captioning infrastructure for image captioning.

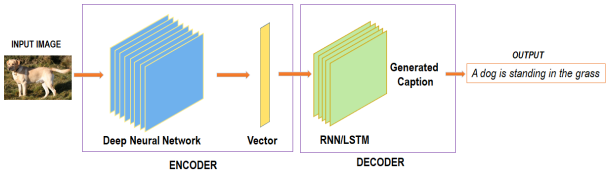

Fig. 1. This is an image captioning model's overall encoder-decoder structure.

As previously stated, our approach is significantly influenced by the work of Xu et al. [31]. Their performance has improved as a result of our efforts. The proposed approach generates a caption encoded as a sequence of encoded words from a single input image. The CNN-Encoder, attention mechanism, and RNN-decoder are the three major components of our proposed framework. They work sequentially, with the images with captions as the CNN-encoder's input. The attention technique and the LSTM work in conjunction to generate captions for the input image, and the output of this is passed to them. The objects and features in the image are retrieved using a convolutional neural network, and then we require a network to construct a meaningful sentence using our information.

$$q = \{q_1, q_2, \ldots, q_c\}, q_i \in R^K \quad (1)$$

A pre-trained CNN as encoder extracts features from images, an attention mechanism weights the image features, and an RNN as decoder provides captions to represent the weighted image features. The overall graphical representation of our framework is shown in Figure 2. K denotes the vocabulary size and c represents the caption length.

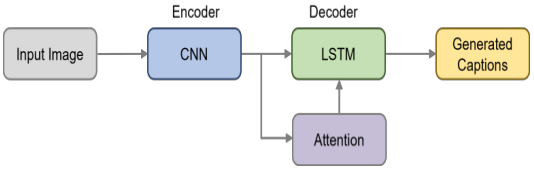

Fig. 2. The graphical representation of the framework.

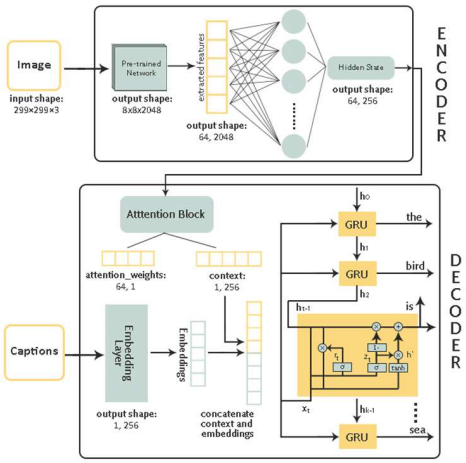

Fig. 3. Figure showing a visual representation of approach. It is based on Xu et al. [31] architecture proposal.

A. Convolutional Neural Network (Encoder)

A CNN pre-trained for an image classification task is commonly used as the encoder in the encoder-decoder framework for image captioning to extract the input image's global representation and sub-region representations. A fully connected layer's output is usually the global representation, while a convolutional layer's output is usually the sub-region representation. As an encoder, we employed a convolutional neural network. A standard feedforward neural network is a convolutional neural network. Each layer performs convolutions and element-wise summations as part of an affine operator. A visual representation is denoted as Figure 3, which illustrates an approach based on the architecture proposal of Xu et al. [31].

$$f(x) = \sum_{k=1}^{K} g_k \times x + b_k \tag{2}$$

We can use the convolutional layers of pre-trained neural networks like Inception, DenseNet169, ResNet101, and VGG16 to convert images into a fixed-size vector. These networks are trained on large datasets and have learned to extract relevant features from images. Here's how we use each of the models mentioned to extract features from images:

- **Inception:** ResNet-101 is a 101-layer deep convolutional neural network. The network learns rich feature representations for a wide range of images because of the many layers. The network's image input size is 224×224 pixels. It generates a $7 \times 7 \times 2048$ feature vector.
- **DenseNet169:**DenseNet-169 is a 169-layer model that has been pre-trained. It has fewer parameters than other approaches, and the architecture effectively tackles the vanishing gradient problem. The network accepts a 224×224 image as input and outputs a $7 \times 7 \times 1664$ feature vector.
- **ResNet101:**ResNet-101 is a 101-layer deep convolutional neural network. The network learns rich feature representations for a wide range of images because of the many layers. The network's image input size is 224×224 pixels. It generates a $7 \times 7 \times 2048$ feature vector.
- **VGG16:**The VGG pre-trained model was released by researchers from the Oxford Visual Geometry Group, who participated in the ILSVRC challenge. By default, the model expects color input photos to be rescaled to 224×224 squares. It generates a $7 \times 7 \times 512$ feature vector.

These networks were trained to classify 1000 different classes of images using the ImageNet dataset. Our goal is not to organize the image but to obtain a fixed-length informative vector for each image. As a result, we eliminated the model's last softmax layer and extracted a fixed-length vector for each image. These extracted characteristics are sent to the RNN, which creates the hidden state using a fully connected layer.

B. Recurrent Neural Networks

While CNNs perform signal processing, they tend to describe patterns in sequences. The outputs of RNNs are fed back into the input, similar to feedforward neural networks. Throughout iterations of this feedback loop, the networks preserve a hidden state that allows them to change their behavior. RNNs are regarded as state-of-the-art in machine translation and other jobs involving text generation because they acquire human grammatical patterns quickly. In this research paper, we use a GRU, a sophisticated variant of the RNN approach that produces a caption by creating one word at each time step based on a context vector, the initial hidden state, and previously made words. The most likely description of an image is obtained in the encoder-decoder approach by maximizing the log-likelihood function of the expression E, considering the related image I and the model parameters θ .

$$\theta^* = \arg\max \sum (I, E) \log p(S \mid I; \theta) \tag{3}$$

Since S it can represent any sentence length, a chain rule is commonly employed to characterize the joint probability over $E_1, E_2, \ldots, E_N$.

$$\log p(E \mid I) = \sum_{t=0}^{N} \log p\left(E_t \mid I, E_0, \ldots, E_{t-1}\right) \tag{4}$$

For clarity, the dependency on θ is excluded. The network training is represented by the pair of (E, I), and we use the *Adam optimizer* to maximize the sum of the log-likelihood functions across the complete training set. A recurrent neural network is used to represent the likelihood $\log p\left(E_{\mathrm{t}} \mid I, E_0, E_1 \cdots, E_{\mathrm{t}-1}\right)$ where there are variable numbers of words that we define up to $t-1$. Using a fixed-sized hidden vector, RNN gives a seamless technique to execute conditioning on prior variables.

$$p\left(E_t \mid I, E_0, E_1, \cdots, E_{t-1}\right) \approx \mathrm{p}\left(\mathrm{E}_t \mid \mathrm{I}, \mathrm{h}_t\right) \tag{5}$$

As a result, at step t, a simple vector replaces the complex conditioning on a variable number of nodes (ht). After the new input X_t, the RNN's hidden state (latent memory) ht is updated with the nonlinear function f.

$$h_{t+1} = f\left(h_t, X_t\right) \tag{6}$$

The capacity of f in Equation (6) to deal with vanishing difficulties and exploding gradients, which are the most typical problems in the development and training of RNN, determines the value of f. Given the inputs X_t, h_{t+1}, the GRU updates for the time step t. To begin, we use the formula to calculate the update gate z_t for the time step t :

When X_t is linked to a network unit, its weight W_z is multiplied. The same is true for h_{t-1}, which stores data from earlier $t-1$ units and is multiplied by their weight W_z. Both results are combined, and the result is squashed between 0 and 1 using a sigmoid activation function. The update gate aids the model in determining how much previous data (from earlier time steps) should be passed on to the future. This is extremely useful since the model can duplicate all of the data from the past, eliminating the risk of disappearing gradients. The framework utilizes the reset gate to determine how much information from the past should be forgotten. We use the following formula to compute it in Equation (7),

$$r_{\mathrm{t}} = \sigma\left(W_{\mathrm{r}} x_{\mathrm{t}} + U_{\mathrm{r}} h_{\mathrm{t}-1}\right) \tag{7}$$

We create new memory content that will store relevant information from the past using the reset gate. It is calculated as follows in Equation (8),

$$h_{\mathrm{t}} = \tanh\left(W x t + r_{\mathrm{t}} \odot U h_{\mathrm{t}-l|}\right) \tag{8}$$

This multiplies the input X_t by weight W and the input h_{t-1} by a weight U. Calculate the Hadamard (element-by-element) product between r_t and Uh_{t-1} . This will determine what time steps should be removed from the preceding ones. Add the results together and use the $tanh$ nonlinear activation function. The network's final step is to calculate the h_t vector, which contains information for the current unit and sends it down to the network. The update gate is required to accomplish this. It determines what should be collected from the present memory content $h't$ and what should be collected from the previous stages h_{t-1}. This is how it's done:

$$h_{\mathrm{t}} = z_{\mathrm{t}} \odot h_{\mathrm{t}-1} + (1 - z_{\mathrm{t}}) \odot h'_{\mathrm{t}} \tag{9}$$

This will execute element-wise multiplication on the updated gates z_t and h_{t-1}, as well as ,$(1 - z_{\mathrm{t}})$, $h't$, and total the results. Below Figure 4 is a visual representation of GRU.
We want the representation of words to be such that the vectors for words with similar meanings or contexts are close to one other in the vector space. The word2vec algorithm, which turns a word into a vector, is likely the most used. A corpus of texts relating to the domain we are engaged in is used to train word2vec.

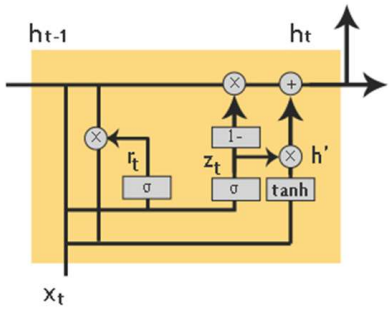

Fig. 4. Visual representation of Gated Recurrent Unit(GRU).

The word2vec algorithm, which turns a word into a vector, is likely the most used. A corpus of texts relating to the domain we are engaged in is used to train word2vec. Word2vec trains a system that can anticipate the surrounding words of a target text to calculate the word vectors. Surrounding words are described as words that appear from both sides of a given word in a small context window of a particular size. These three sentences will serve as our corpus. (1) *This research paper is about deep learning and computer vision.* (2) *We love deep learning.* (3) *We love computer vision.* The challenge is to predict the context terms *"learning,"* *"and,"* and *"vision"* from the first phrase and *"we,"* *"love,"* and *"vision"* from the last sentence given the word *"computer."* As a result, training aims to maximize the log probabilities of these context terms given the word *"computer."* The formulation is in Equation (10),

$$\text{Objective } = \max \sum_{\mathrm{t}=1-\mathrm{m} \geq \mathrm{j} \geq \mathrm{m}}^{T} \sum_{\mathrm{w}} \log \mathrm{P}\left(\mathrm{E}_{\mathrm{wt}+\mathrm{j}} \mid \mathrm{E}_{\mathrm{wt}}\right) \tag{10}$$

Where m is the size of the context window and t is the length of a corpus. The similarities or inner product between the context word vector and the center word vector represents $P\left(E_{wt+j} \mid E_{wt}\right)$ when a word appears in context. When it is the central word, it has two vectors associated with it, denoted by R and S, accordingly. As a result, $P\left(E_{wt+j} \mid E_{wt}\right)$ it is defined as:

$$\frac{\mathrm{e}^{R_{Wt+j}^{T}}}{\sum_{i=1}^{M} e^{R_{i}^{T} S_{Wt}}} \tag{11}$$

The denominator is the normalization term, which compares the similarity of the central word vector to context vectors of every other word in the lexicon, resulting in a probability of one. The vector representation for a word is then chosen as the center vector, as with GRU.

C. Attention models

We apply the Bahdanau attention mechanism to isolate the image content, which has been widely used to tackle the challenge of image categorization since it eliminates the need to process every pixel in an image. Instead of taking features from the entire image, the salient portion of the image is determined at each step and input into the RNN. The algorithm uses the image to create a focused view and predicts the relevant term to that location. The location where attention is directed must be determined based on previously generated words. Otherwise, new words formed within the region may be coherent, although not in the description generated. After that, proceed to the mechanism proposed by Bahadano. To begin, calculate the e_{jt} score in Equation (12),

$$e_{jt} = f_{\mathrm{ATT}}\left(E_{t-1}, h_j\right) \tag{12}$$

e_{jt} is a score that indicates how essential an image's jth pixel is at every time step t of the decoder. The prior state of the decoder is S_{t-1}, while the current state of the encoder is h_j . f_ATT Is a primary feed-forward neural network that sums S_{t-1} and h_j from the fully connected layer, then passes it through a nonlinear function $tanh$ before returning to the fully connected layer.

$$\mathrm{e_{jt}} = \mathrm{FC}\left(\tanh\left(\mathrm{FC}\left(\mathrm{E_{t-1}}\right) + \mathrm{FC}\left(\mathrm{h_i}\right)\right)\right) \tag{13}$$

We use softmax to get the probability distribution in Equation (14),

$$\alpha_{jt} = \mathrm{softmax}\left(e_{jt},\ \mathrm{axis}\ = 1\right) \tag{14}$$

Softmax is usually applied to the last axis; however, since the shape of the score is (batch-size, max-length, hidden-size), we want to apply it to the 1^{st} axis. The maximum length of our input is max length. Softmax should be applied to that axis because we aim to allocate weight to each input. Now that we have the input, we must feed the decoder a weighted sum combination of the input.

$$c_t = \sum_{j=1}^{T} a_{jt} h_j \ \text{suchthat} \ \sum_{j=1}^{T_x} a_{jt} = 1 \ \text{and} \ a_{ij} \geq 0 \tag{15}$$

The context vector (weighted total of the input) that will be sent to RNN isc_t.

$$\mathrm{E_t} = \mathrm{RNN}\left(\mathrm{S_{t-1}}, \mathrm{e}\left(\mathrm{y'_{t-1}}\right), \mathrm{c_t}\right) \tag{16}$$

The previous state of the decoder is S_{t-1}, and the last predicted word is $e\left(y'_{t-1}\right)$.

IV. Test Configuration and Results

To assess the effectiveness of the proposed GRU-based attention model, we conducted a comprehensive set of experiments using the popular MSCOCO and Filckr30K datasets. This section presents the experimental results obtained by applying our model to various pre-trained CNN architectures: InceptionV3, DenseNet-169, ResNet101, and VGG16. We also compare the performance of our model against state-of-the-art approaches to showcase its competitive performance. To evaluate the framework's effectiveness, metrics such as BLEU, ROUGE, CIDER, and METEOR were used to compare the generated captions to the ground truth captions. These metrics measure the similarity between the generated and ground truth captions based on n-gram overlap, recall, and other factors.

A. Datasets

MSCOCO: MSCOCO dataset [9], which contains $82,783$, $40,504$, and $40,775$ images for training, validation, and testing, is the most significant benchmark dataset for the image captioning task. Because most images feature many objects in the context of complicated situations, this dataset is difficult to analyze. Each image in this dataset has five captions with different ground truths annotated by humans, as seen in 4. By utilizing the same data split as in [23] for offline assessment, which consists of $5,000$ photos for validation, $5,000$ images for the test, and $113,287$ images for training. After that, combine the testing set with the training set to create a more extensive training set for online evaluation on the MSCOCO evaluation server. Then, remove all non-alphabetic characters from the captions, convert all letters to lowercase, and tokenize the captions with white space. As a result, a vocabulary of $9,487$ terms has been created.

Fig. 5. Representation of dataset.

Flickr30K: Flickr30K is an automated image interpretation and grounded language comprehension dataset. It comprises 30K Flickr images and has 158K captions generated by personal annotators. It doesn't have a specified division of images for analysis and justification of instruction. Investigators may select their choice numbers for preparation, measurement, and evaluation. The dataset also includes typical object detectors, a color classifier, and a tendency against more significant object collection. The Flickr30k dataset contains $31,000$ images, each with five captions. Table 1 compares the splitting ratios of various datasets briefly.

TABLE I
SPLITTING RATIOS OF MSCOCO AND FLICKER30K DATASETS.

Dataset	Train Split	Validation Split	Test Split
MSCOCO	113,287	5,000	40,504
Flicker30k	29,000	1,014	1,000

B. Evaluation measurements

Four commonly used evaluation metrics, namely BLEU 1-4 (Bilingual Evaluation Understudy) [32], Meteor (Metric for Evaluation of Translation with Explicit Ordering) [33], Rouge-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence) [34], and CIDEr (Consensus-based Image Description Evaluation) [35], are utilized to evaluate the quality of generated sentences utilizing the publicly accessible MSCOCO tools to analyze the performance of our proposed methods quantitatively. These metrics compare the consistency of n-grams in generated and reference sentences. To make accurate comparisons with existing image captioning approaches.

C. Quantitative Results

Demonstrating the quantitative results in this section shows the efficacy of the proposed strategy. The suggested technique is compared to seven state-of-the-art models on MSCOCO and Flickr30k in a multi-comparison, as shown in Tables 2 and 4, respectively.

TABLE II
COMPARISON OF THE PROPOSED APPROACH WITH STATE-OF-THE-ART METHODS ON THE MSCOCO DATASET.

MODE	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
Google NIC [36]	0.67	0.45	0.30	0.20	–	–	–
Soft Att [31]	0.71	0.49	0.34	0.24	–	–	0.24
MSM [2]	0.73	0.57	.043	0.33	0.54	1.02	0.25
Attribute-driven Att [37]	0.74	0.56	0.44	–	0.55	1.104	–
NBT [38]	0.75	–	0.34	–	–	1.107	0.27
Context-aware att [39]	0.76	0.60	0.46	0.36	0.56	1.103	0.28
GCN-LSTM [40]	0.77	–	–	0.36	0.57	1.107	0.28

In the first step of the LSTM-based language model, NIC injects image features derived from the fully connected layer of a deep CNN. The results presented are cited directly [36]. Soft-Att selects some regional representations from the deep CNN's final convolutional layer and uses an LSTM-based language model to decode each word at each time step, depending on the representations selected [31]. MSM incorporates inter-attribute correlations into a multiple-instance learning approach and investigates several techniques of injecting detected characteristics and image representations into an LSTM-based language framework [2]. Attribute-driven using CNN-RNN architecture and the visual attention method for attribute detector, attention model the co-occurrence dependencies among attributes [37]. NBT architecture for visually grounded image captioning generates free-form natural language descriptions while localizing things in the image [38]. The relationship between the objects/regions in an image is modeled with a GNN in visual context-aware attention, which considers the visual relationship between areas of interest for improved representation of the visual content in the image [39]. The elegant views of what kind of visual relationships could be built between objects and how to nicely leverage such visual relationships to learn more informative and relation-aware region representations come from GCN-use LSTMs of visual relationships for enriching region-level models and eventually enhance and lead to the elegant views of what kind of visual relationships could be built between objects and how to nicely leverage such visual relationships to learn more informative and relation-aware region representations [40]. The bold number represents the top results for that measure, whereas those with a dash (–) are unavailable.

TABLE III
SUMMARIZED OUTCOMES OF THE PROPOSED FRAMEWORK ON THE MSCOCO DATASET.

MODE	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
Inception V3	**0.78**	**0.57**	**0.44**	0.36	**0.59**	1.105	0.27
VGG16	0.74	**0.57**	**0.44**	0.33	0.56	1.109	0.26
DenseNet169	0.74	0.56	0.43	0.36	0.58	1.103	0.27
ResNet101	0.75	0.56	**0.44**	**0.37**	**0.59**	1.104	**0.29**

TABLE IV
COMPARISON OF THE PROPOSED APPROACH WITH STATE-OF-THE-ART METHODS ON THE FILKER30K DATASET.

MODE	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
Google-NIC [36]	0.66	0.42	0.27	0.18	–	–	–
m-GRU [41]	0.66	0.40	0.28	0.20	0.29	0.48	–
phi-LSTM [42]	0.64	0.45	0.31	0.21	0.19	0.45	0.44
Soft-Att [31]	0.66	0.43	0.28	0.19	0.18	–	–

To compare the performance of our proposed model with other state-of-the-art approaches, we utilized different pre-trained models for encoder architecture by evaluating the quantitative analysis of the presented framework's four metrics: BLEU, Rouge, CIDER, and Meteor. Table 3 shows the results for the MSCOCO dataset, while Table 5 displays the results for the Fliker30k dataset. The top performers for each metric are indicated by **bold** numbers. Inception outperformed the other models, followed by ResNet10. While the overall results are satisfactory, we acknowledge using 113,287 images for training from the MSCOCO dataset and 29,000 from the Filker30k dataset. Figure 5 illustrates the loss plot for four pre-trained networks. This could be useful in evaluating the training process and comparing the performance of different pre-trained networks. Figures 6 and 7 dispute the visual representation for the performance of proposed GRU attention-based models.

TABLE V
SUMMARIZED OUTCOMES OF THE PROPOSED FRAMEWORK ON FLICKR30K DATASET.

MODE	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
Inception V3	**0.70**	**0.47**	0.36	0.26	0.47	0.49	0.28
VGG16	0.67	0.46	0.36	0.24	0.44	**0.47**	**0.29**
DenseNet169	0.65	0.45	0.34	**0.27**	0.46	**0.47**	0.27
ResNet101	0.67	0.46	**0.36**	0.26	0.46	0.46	0.28

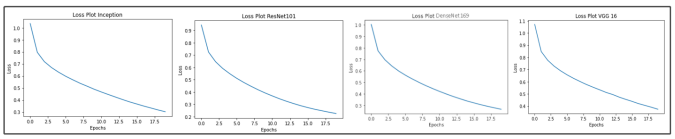

Fig. 6. Loss Plot of models pre-trained with different networks.

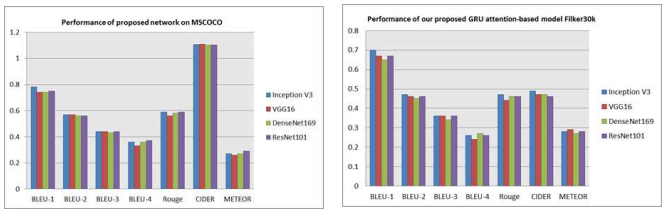

Fig. 7. (a) Statistical evaluation score of different models on MSCOCO dataset. (b) Statistical evaluation score of different models on the Filker30K dataset.

D. Qualitative results

The qualitative results of the proposed model are intriguing as it generates relevant and grammatically correct captions for various images. Figure 8 demonstrates positive outcomes on the MSCOCO dataset, while Figure 9 illustrates the caption generation on the Flickr30k dataset.

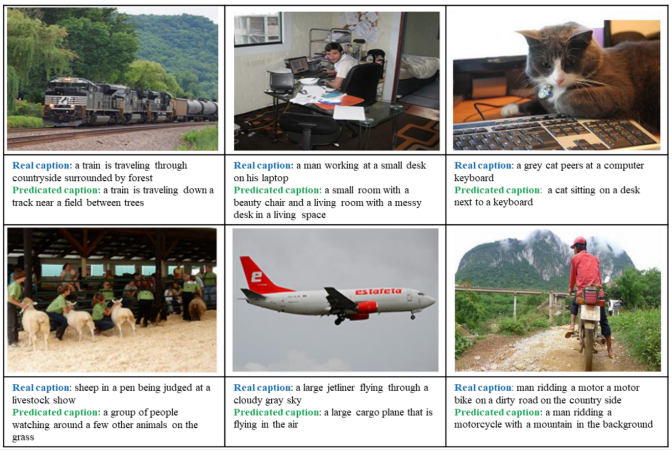

Fig. 8. Example from the conventional approach with the proposed framework and their ground truth captions for the MSCOCO.

Although most captions are informative, there are instances where the generated caption describes a situation different from the image or is entirely incomprehensible. These errors can be attributed to the system's two tasks: image recognition and text generation. The former can be labeled as a failure in image recognition, while the latter can be attributed to a failure in text generation.

The study conducted relevant test experiments to validate the effectiveness of a newly developed framework for automatic caption synthesis. The framework demonstrated improved results in generating captions for images not part of the training or validation datasets. The research indicated that the framework excelled in accurately captioning various images.

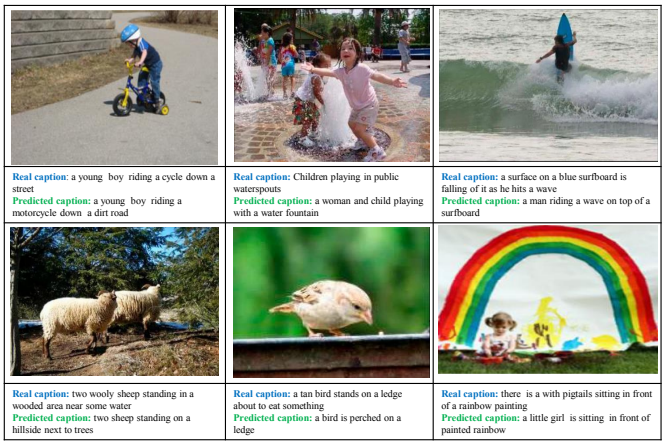

Fig. 9. Image-to-text retrieval results.from the conventional approach with the proposed model for the Filker30K dataset.

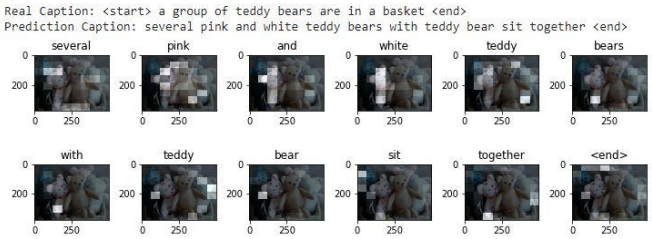

Fig. 10. Attention is drawn to different regions of the image. The attention weights in white spots are higher

However, the network occasionally encountered challenges in distinguishing tasks, resulting in failures in both image recognition and text generation. Despite these failures, the attention network architecture displayed promising outcomes, as showcased in Figure 10, where the network's focus was drawn to relevant objects, such as *teddy bears*, resulting in informative captions.

V. Conclusions and Future Scope

In this research, we proposed a single joint model for automatic image captioning using a combination of CNN and GRU with an attention network. The suggested framework outperforms pre-trained CNNs and significantly improves caption quality, as demonstrated by both qualitative and quantitative experiments using the MSCOCO and Flickr30k datasets t. Presented findings show that the Bahdanau attention model combined with GRU can effectively focus on specific regions of the image and enhance the model's overall performance. The proposed approach can help bridge the gap between computer vision and natural language processing and extend caption generation into specific domains. Future research can explore various directions to enhance the performance of image captioning models. The first possible direction is to investigate the use of more advanced attention mechanisms, such as the Transformer-based models, to improve further the

model's ability to focus on specific image regions. Furthermore, explore the use of alternative language models, such as the LSTM, to assess their impact on caption quality. Additionally, incorporating external knowledge sources, such as common sense reasoning, could be beneficial to generate captions that are more informative and accurate. Finally, including a feedback loop into the model to iteratively refine the captions could be explored as a promising direction for future research.

Acknowledgment:
The Fundamental Research Funds support this research for the Central Universities.(Grant no.WK2350000002).

REFERENCES

[1] M. Pedersoli, T. Lucas, C. Schmid, and J. Verbeek, "Areas of Attention for Image Captioning," 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1251-1259

[2] Z. Wu, T. Yao, Y. Fu, and Y.-G. Jiang, "Deep learning for video classification and captioning," Frontiers of multimedia research, pp. 3-29, 2017

[3] R. Khan, M. S. Islam, K. Kanwal, M. Iqbal, M. I. Hossain, and Z. Ye, "Attention-based sequence-to-sequence framework for auto image caption generation," Journal of Intelligent and Fuzzy Systems, vol. 43, pp. 159-170, 2022.

[4] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7008-7024. 2017

[5] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering." pp. 6077-6086, 2018

[6] C. Chunseong Park, B. Kim, and G. Kim, "Attend to you: Personalized image captioning with context sequence memory networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 895-903.

[7] H. Zhu, R. Wang, and X. Zhang, "Image captioning with dense fusion connection and improved stacked attention module," Neural Processing Letters, vol. 53, no. 2, pp. 1101-1118, 2021.

[8] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," Advances in neural information processing systems, vol. 29, 2016. 2361-2369.

[9] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context." pp. 740-755, 2014.

[10] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3156-3164. 2015.

[11] G. Huang and H. Hu, "C-Rnn: a fine-grained language model for image captioning," Neural Processing Letters, vol. 49, no. 2, pp. 683-691, 2019.

[12] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3128-3137, 2015. ISBN 978-1-4673-6964-0

[13] P. Cao, Z. Yang, L. Sun, Y. Liang, M. Q. Yang, and R. Guan, "Image captioning with bidirectional semantic attention-based guiding of long short-term memory," Neural Processing Letters, vol. 50, no. 1, pp. 103-119, 2019.

[14] L. Yang and H. Hu, "Adaptive syncretic attention for constrained image captioning," Neural Processing Letters, vol. 50, no. 1, pp. 549-564, 2019

[15] Z. Ye, R. Khan, N. Naqvi, M.S. Islam, A novel automatic image caption generation using bidirectional long-short term memory framework, Multimedia Tools and Applications, 80 (2021) 25557-25582

[16] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651-4659.

[17] Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, "Deep reinforcement learning-based image captioning with embedding reward," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 290-298.

[18] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 7008-7024.

[19] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in International conference on machine learning, 2014: PMLR, pp. 595-603.

[20] E. Arican, T. Aydin, An RGB-D descriptor for object classification, Romanian Journal of Information Science and Technology (ROMJIST), 25 (2022) pp. 338-349..

[21] A. Verma, T. Meenpal, B. Acharya, Computational cost reduction of convolution neural networks by insignificant filter removal, Science and Technology, 25 (2022) pp. 150-165.

[22] D. Borlea, R.-E. Precup, A.-B. Borlea, Improvement of K-means cluster quality by post processing resulted clusters, Procedia Computer Science, 199 (2022) pp. 63-70.

[23] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3128-3137.

[24] C. He and H. Hu, "Image captioning with text-based visual attention," Neural Processing Letters, vol. 49, no. 1, pp. 177-185, 2019.

[25] C. Liu, J. Mao, F. Sha, and A. Yuille, "Attention correctness in neural image captioning," in Thirty-first AAAI conference on artificial intelligence, 2017.

[26] Z. Yang, Y. Yuan, Y. Wu, W. W. Cohen, and R. R. Salakhutdinov, "Review networks for caption generation," Advances in neural information processing systems, vol. 29, 2016

[27] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651-4659

[28] L. Zheng, Z. Caiming, and C. Caixian, "MMDF-LDA: An improved Multimodal Latent Dirichlet Allocation model for social image annotation," Expert Systems with Applications, vol. 104, pp. 168-184, 2018

[29] C. Sur, "aiTPR: attribute interaction-tensor product representation for image caption," Neural Processing Letters, vol. 53, no. 2, pp. 1229-1251, 2021

[30] A. Adhikari and S. Ghimire, "Nepali image captioning," in 2019 Artificial Intelligence for Transforming Business and Society (AITB), 2019, vol. 1: IEEE, pp. 1-6

[31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention." 2015, pp. 2048-2057

[32] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311-318.

[33] A. Lavie and M. J. Denkowski, "The METEOR metric for automatic evaluation of machine translation," Machine translation, vol. 23, no. 2, pp. 105-115, 2009

[34] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74-81

[35] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566-4575,

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156-3164,

[37] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Show, Observe and Tell: Attribute-driven Attention Model for Image Captioning," in IJCAI, 2018, pp. 606-612

[38] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7219-7228

[39] 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 Recognition, vol. 98, p. 107075, 2020.

[40] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 684-699

[41] Li, X., Yuan, A. and Lu, X., 2018. Multimodal gated recurrent units for image description. Multimedia Tools and Applications, 77, pp.29847-29869.

[42] Tan, Y.H. and Chan, C.S., 2019. Phrase-based image caption generator with hierarchical LSTM network. Neurocomputing, 333, pp.86-100.