2024 4th International Conference on Computer Communication and Artificial Intelligence (CCAI) | 979-8-3503-6276-3/24/$31.00 ©2024 IEEE | DOI: 10.1109/CCAI61966.2024.10602893

Visual Insight: Deep Multilayer Fusion with Inception-based LSTM for Descriptive Image Captioning

Rashid Khan
College of Big Data and Internet,
Shenzhen Technology University, Shenzhen, 518188, China.
College of Applied Sciences, Shenzhen University,
Shenzhen, 518060, China.
Email: rashidkhan@sztu.edu.cn

Bingding Huang*
College of Big Data and Internet.
Shenzhen Technology University, Shenzhen, 518188
Shenzhen, China
Email: huangbingding@sztu.edu.cn
*Corresponding author

***Abstract*—The image caption is a technology that aids us in comprehending the contents while employing machines to create descriptive text for an image. The captions are generated using Natural Language Processing (NLP) and Computer Vision (CV). When the descriptions contain a single word like "boy," "cycle," etc., the image captioning work is completed by combining the detection method with image captioning when one predicted region covers the entire image, such as a boy riding a bicycle. to combine the tasks of localization and description It is presently a current hot trend in deep learning development to use it to analyze visual information and write descriptive text. This paper presented a multilayer dense focus image captioning model. We used transfer learning techniques to adjust pre-trained image classification models and integrate them with long short-term memory network (LSTM) architectures to evaluate the performance of each of the combined frameworks. The variable length input is encoded into a fixed-dimensional vector, which is taken as the maximum length of the caption available mapped with the image, and the recurrent neural network (RNN) uses this representation to "decode" it to the desired output sentence. We experimented with the Flickr8k, Flickr30k, VizWiz, and MSCOCO datasets. According to the analysis of experimental data on evaluation criteria, the model described in this research can effectively accomplish image captions according to the analysis of experimental data. Its performance is better than classic image captioning algorithms.**

* ***Index Terms*—Image captioning, CNN, Inception V3, LSTM, Natural Language Processing.**

I. Introduction

Billions of images are taken every day. It is a challenging task to classify and organize them in a way that allows us to recover a specific group of pictures or a unique image quickly and easily. or AI development and, particularly, how this technology helps teach machines to recognize objects in visual media. Search engines widely use automatic image captioning to retrieve and show relevant search results to the user over the annotation keywords, to categorize personal multimedia collections, for automatic product tagging in online catalogs, computer vision development, and other business and research areas [1], [2]. In Artificial Intelligence, image captioning is a fundamental goal that explains objects, attributes, and relationships in an image in a natural language sense [3]. When people share photographs on social media, they frequently use these descriptions. Lastly, perceptual reports describe the low-level visual features of images, such as whether the image is a painting or a drawing or the dominant colors or patterns [4]. It inputs an image and outputs a short textual summary of the photo's content [5]. The machine must read and transform an array of numbers into a meaningful sequence of words, as seen in Figure I.

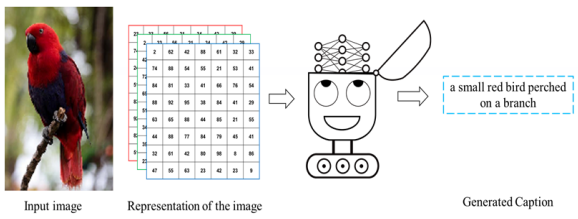

Fig. 1. An example of a machine-generated natural language description for a given image.

For decades, the computer vision community has been working towards the ultimate objective of complete image understanding. It has several applications, including semantic image search, supporting chatbots with visual insight, and allowing people with visual impairments to see the world around them [6], [7]. The problems related to image caption generation have emerged as a greater spectrum of research in NLP and computer vision [8], [9]. Understanding an image's salient and semantic content requires a machine [10], the object type, and the relationship between those objects, etc. The method is significant for human life because it directs visually disabled people to understand the visualized environment by organ or to create an actual sense of intellectual robot that exposes language, sensory, and visual networks [11]. A well-known encoder-decoder scheme, in which an RNN is deployed to encode the source sentence [12], and an additional RNN is deployed to predict the target sentence,

*Corresponding author: Bingding Huang (huangbingding@sztu.edu.cn)

is currently working on this topic [13]. This method has established its appliances in several fields: text recognition [14] and speech recognition [15]. It is important to classify the leading methods into two main streams. An end-to-end, encoder-decoder architecture implemented from machine translation takes one stream. A convolutional neural network (CNN) [12], for instance, was used to obtain high-level image attributes, and then they were fed into an LSTM for caption generation. Search engines widely use automatic image captioning to retrieve and show relevant search results to the user over the annotation keywords, to categorize personal multimedia collections, for automatic product tagging in online catalogs, computer vision development, and other business and research areas. We've picked some interesting use cases where this technology can be helpful and profitable. this neural system for image captioning uses an image as input, and the output is a sentence describing the visual content of a picture. Image captioning can be regarded as an end-to-end Sequence problem, as it converts images, which are regarded as a sequence of pixels, to a sequence of words [23]. For this purpose, we need to process both the language or statements and the images. We use recurrent neural networks for the language part, and for the image part, we use CNN to obtain the feature vectors. This paper uses deep learning to design an image captioning model based on the encoder-decoder structure. The extended deep CNN is used as an encoder to extract image features, and the -LSTM network is used to generate descriptive sentences. This paper focuses on the end-to-end automatically generated image captioning model, and the main contributions are as follows:

The main contributions of our work are as follows:

- Using transfer learning to adopt a classification model and perform feature extraction on images
- We innovatively employed transfer learning techniques to adapt pre-trained image classification models and seamlessly integrate them with LSTM architectures. This approach facilitated efficient feature extraction from images, enhancing the model's ability to generate accurate and contextually relevant captions.
- A hybrid deep neural framework for image caption generation, leveraging the powerful InceptionV3 model for feature extraction. By implementing a fully integrated image detection model, we achieved superior performance on benchmark datasets like Flickr8k, Flickr30k, VizWiz [17], and MSCOCO [16], surpassing baseline models across various evaluation metrics.
- We thoroughly evaluated our proposed model, utilizing automatic evaluation metrics such as BLEU, METEOR, and CIDEr. Through rigorous comparison with existing image captioning algorithms, our model consistently demonstrated superior performance, highlighting its effectiveness in generating high-quality image captions.

II. Related Work

The strategies of image captioning can be categorized into three groups [18]. template-based approach [19], [20] creates captions based on templates for languages. The scan-based approach [21] checks a sentence pool for the most semantically related captions [12]. Current research primarily focuses on language-based strategies with an encoder-decoder framework [13], [22]–[24], where a CNN encodes images into visual attributes and decodes attributes into sentences by an LSTM [15], [25]. It has been shown that mechanisms of attention [3], [23] and high-level attributes and concepts will assist with captioning generation. Two other approaches are used to boost the efficiency of sequencing learning models [23]. A new attention technique, the so-called TA-LSTM [21], was initially implemented to manage the image background information at each level of LSTM. The LSTM structure was subsequently redesigned, where the stack LSTM and the parallel LSTM were deployed to achieve better performance over the regular LSTM. CNN has been implemented in [15], whereby the optimum title is chosen and accuracy is increased. But it could not reliably count the objects. Furthermore, the LSTM model is deployed in and improves visual attention. However, the use of similitude between graphs requires consideration. Similarly, gLSTM was used in, offering increased sensitivity and balancing semantic information. Nevertheless, inaccurate explanations may be made. The implementation of CNN provides improved accuracy and an ideal title, but more needs to be done with the dense image caption. In [37], VD-SAN was suggested for better recognition of attributes and to improve the captioning framework. However, it may confuse the LSTM caption generator.

III. Model Architecture

As we have observed using the traditional CNN-RNN model, a vanishing gradient problem hinders the Recurrent Neural Network from learning and getting efficiently trained. So, to reduce this gradient descent problem, we propose this model in this study to increase the efficiency and accuracy of generating captions for the image. Figure 2, given below, is the architecture for our proposed model.

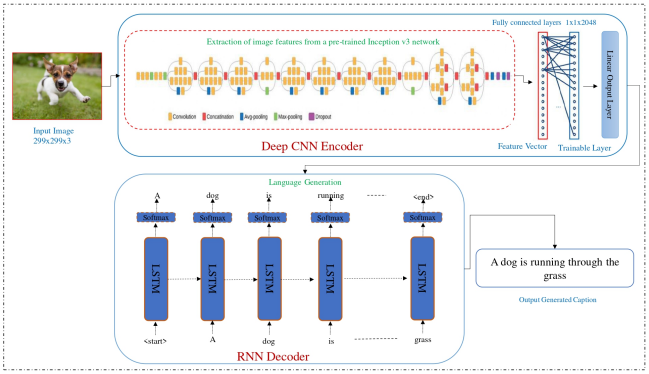

Fig. 2. Architecture of Deep CNN-RNN Model.

We implemented a deep recurrent architecture that can generate an English sentence that describes objects, actions, or events in an RGB image:

$$F = q(M) \tag{1}$$

where M is an RGB image and F is a sentence, and q is the function that need to learn. We used Flickr8k and MSCOCO, Flickr8k, and VizWiz datasets. The captions are gathered from human beings using Amazon's Mechanical Turk (AMT). We manually checked some examples side-by-side, comparing the image and corresponding sentences. We found the captions to be very expressive and diverse. Our models use a CNN, which was pre-trained on Inception v3, to obtain image features. We then feed these features into a vanilla RNN or an LSTM network to generate an English image description.

IV. Methodologies adopted for Caption Generation

To feed data as input to the neural network, every image in the dataset needs to be converted into fixed-size vectors. The given images were scaled down to 299 × 299, the input shape expected by the Inception Network. The preprocessing function was applied to the given images. The captions were read, and the genism library was used to preprocess them. All captions were converted to lowercase; the genism library removed numbers, symbols, and punctuations. All words of length one were removed. The given captions were padded with zeros, considering a maximum length of 35. Start and End indicators were added to all the captions. Only those words considered in the vocabulary appear at least 10 times in the data. This was done to reduce model underfitting and reduce data size for training. Thus, after applying global average pooling, the model's output would be a tensor of shape (1,2048). Feeding a 299 × 299 RGB image to this model encodes the image to the data cleaning functions. For image caption generation, LRCN maximizes the probability of the description giving the image:

$$\theta^* = \arg\max_{\theta} \sum_{(M,P)} \log p(F \mid M; \theta) \tag{2}$$

where F is the input image, and F is the input sample sentence. Let the length of a sentence N. The method utilizes the chain rule to model the joint possibility in the below equation $F_0, ..., F_{c-1}$.

$$\log p(F \mid M) = \sum_{c=0}^{N} \log p\left(F_c \mid M, F_0, \cdots, F_{c-1}\right) \tag{3}$$

where $theta$ is omitted for ease, F_c represents the word at time c. The model consists of two components. The initial component is a CNN that converts the image into a visual feature of fixed length. The input e is incorporated as the visual attribute in the RNN.

$$v = R_e(\mathbb{CNN}(M)) \tag{4}$$

Wv represents the visual feature embedding. The RNN has a consistent visual feature at every step. In the RNN, every word is symbolized by a one-hot vector F_c with a dimension matching the size of the dictionary. Special start and stop words are represented by F_0 and F_N. R_F is the parameter for word embedding.

$$x_c = R_c F_c, c \in \{0 \cdots Z-1\} \tag{5}$$

This means the image and words are assigned to identical coordinates in a single space. Following the RNN's internal calculations, the characteristics v, x_c, and internal hidden variable h_c are translated into a likelihood to forecast the word at the present moment.

$$p_{c+1} = \mathrm{LSTM}\left(e, x_c, h_c\right), c \in \{0 \cdots Z-1\} \tag{6}$$

A sentence with a higher likelihood does not guarantee its correctness compared to other possible sentences, so a technique like *Beam Search* produces additional sentences and selects the *top-K* sentences.

A. Inception-v3 for Feature Extraction

Inception-v3 is a classical deep network composed of 11 Inception modules of five kinds in total. Experts with a convolutional layer, activation layer, pooling layer, and batch normalization layer design each module. These modules are concatenated to achieve maximum feature extraction in this Inception-v3 model. Here, the Inception modules adopt the multi-scale concept. Every module includes numerous branches with diverse kernel sizes like (1 × 1, 3 × 3, 5 × 5 and 7 × 7). These filters extract and concatenate diverse scales of feature maps and transmit the combinations to the subsequent phase. In every inception module, 1×1 convolutions are deployed for reducing dimensions before deploying "computationally expensive 3×3 and 5×5 convolutions". RGB image to this model encodes the image. Figure 3 visually outlines the architecture, clearly understanding the module's components and their interactions.

Inception v3 starts with convolutional layers, which perform feature extraction from the input image. These layers apply convolution operations, often followed by batch normalization and activation functions such as ReLU. The output of a convolutional layer l can be represented as:

$$\mathrm{O_l} = \mathrm{ReLU}\left(\mathrm{W_1} \times \mathrm{I_{l-1}} + \mathrm{b_l}\right) \tag{7}$$

where W_l represents the weights $I_(l-1)$ is the input to the layer, b_l is the bias term, and x denotes the convolution operation. The core of Inception v3 architecture consists of multiple Inception modules. These modules use a combination of convolutional filters of different sizes to capture information at different scales. Towards the end of the network, there are fully connected layers that perform high-level feature aggregation and classification.

$$O_{\mathrm{fc}} = \mathrm{ReLU}\left(W_{\mathrm{fc}} \cdot O_{\mathrm{flatten}} + b_{\mathrm{fc}}\right) \tag{8}$$

where the weights are represented, $O_flatten$ is the flattened output of the previous layer. The final layer of Inception v3 typically uses a softmax activation function to produce class probabilities. The output of the softmax layer can be represented as:

$$\text{softmax}\,(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{9}$$

z_i represents the input to the softmax function for class and is the total number of classes. Inception-v3 is a classical deep network composed of 11 Inception modules of five kinds in total. Experts with a convolutional layer, activation layer, pooling layer, and batch normalization layer design each module. These modules are concatenated to achieve maximum feature extraction in this Inception-v3 model. Here, the Inception modules adopt the multi-scale concept. Every module includes numerous branches with diverse kernel sizes. These filters extract and concatenate diverse scales of feature maps and transmit the combinations to the subsequent phase. In every inception module, 11 convolutions are deployed to reduce dimensions before deploying "computationally expensive convolutions." RGB image to this model encodes the image.

$$\text{CNN Encoder: } B = Z \times q \tag{10}$$

Equation 10 describes how the CNN model works where input, q = filter, $\times$ is a convolution operation, and is the extracted feature. A filter is an array of values with specific values for feature detection. The filter moves to every part of the image and returns a high value if the feature is detected. If the feature is not present, then a low value is returned. The filter can be described as a kernel, a matrix with specific values. The image is also a matrix with different values corresponding to the pixel's intensity at that point ranging from 0–255. This helps in extracting features from the image. A 33 size kernel was adopted experimentally for the model used in this research, as it offered greater accuracy in the time required to traverse the whole scan.

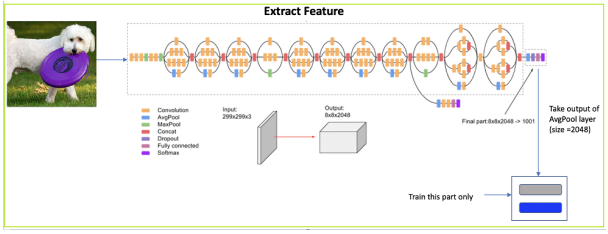

Fig. 3. Architecture of inception v3 module.

B. Language Decoder

LSTM models are used in the domain of NLP. This part of the model concatenates the CNN and LSTM parts. Given an input sequence, a basic RNN model generates the output sequence, which depends on the input length. Between the input layer and the output layer, there is a hidden layer, and the currently hidden state ht is estimated using a recurrent unit:

$$h_t = f\,(h_{t-1}, x_t) \tag{11}$$

where is the current input, h_{t-1} is the previous hidden state, and f can be an activation function or other unit accepting both as input and producing the current output ht. The concatenated data is fed to a 256-unit dense layer followed by a dropout layer to prevent overfitting. Another dense layer has been added with the softmax activation for the inferencing part. The number of units in this layer has been set equal to the vocabulary size. This outputs a sequence of the next probable words to create a generated caption. LSTM layers can also be added to this part of the model for more complex models. In Figure 4, the LSTM cell indicates that activation functions are hidden layer vectors and input vectors.

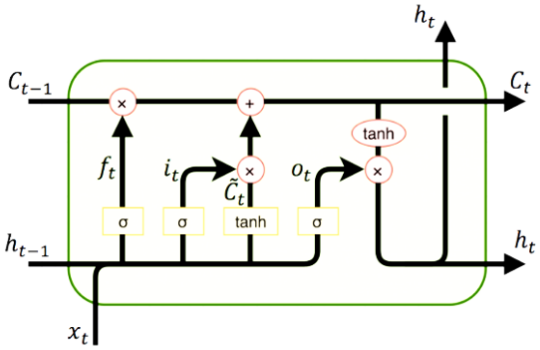

Fig. 4. Long short-term memory unit.

$$i_t = \sigma\left(U_x^i x_t + U_h^i h_{t-1} + b_i\right) \tag{12}$$

$$f_t = \sigma\left(U_x^f x_t + U_h^f h_{t-1} + b_f\right) \tag{13}$$

$$o_t = \sigma\left(U_x^o x_t + U_h^o h_{t-1} + b_o\right) \tag{14}$$

$$\widetilde{c_t} = \tanh\left(U_x^c x_t + U_h^c h_{t-1} + b_c\right) \tag{15}$$

$$c_t = f_t \times c_{t-1} + i_t \times \widetilde{c_t} \tag{16}$$

$$h_t = o_t \times \tanh\left(c_t\right) \tag{17}$$

where $U^i., U^f., U^0$. are weight matrices for corresponding gates, and b. are bias terms learned from the network. The input gate i decides the degree to which new memory is added to the memory cell. The forget gate f determines the degree to which the existing memory is forgotten. The memory cell c is updated by forgetting part of the existing memory and adding new memory.

C. Sentence generation

The output of LSTM is the probability of each word in the vocabulary. Beam search is used to generate sentences. Beam search is a heuristic search algorithm that explores a graph by expanding the most promising node in a limited set. In addition to beam search, we use k best search to generate sentences. It is very similar to the time-synchronous Viterbi search. The

method iteratively selects the k best sentences from all the candidate sentences up to time **t** and keeps only the resulting best **k** of them.

V. Experimental Setup

To design a well-implemented Image Caption Generator experiment, we had to carefully plan it out to ensure we could provide the most accurate and encompassing captions for our images. For this reason, we chose three publicly available datasets, Flickr8k, Flickr30k, and MSCOCO, each containing images and their respective captions, which can be used for training and testing purposes. We used transfer learning for the InceptionV3 architecture since we used pretrained weights to extract features of our images, and for the language model, we used a pretrained word embedding model to improve it. Captions were pre-processed rigidly through tokenization and vocabulary constriction to maximize the model efficiency. Training data were prepared as triplets of (image, source caption, and target caption), where the LSTM cell predicts the target caption words from image representations and previously generated words. The reference architecture of the model was an Encoder-Decoder architecture, with intermediate model weights saved for later re-training. Different evaluation metrics were used to assess the quality of the generated caption. Text normalization and convolution architecture with state-of-the-art technology were used to maximize the model's efficiency. In conclusion, the experimental setup was designed and conducted to maximize the effectiveness and reliability of the Image Caption Generator with diverse image datasets.

A. Datasets and Evaluation Measurements

This section provides insight into the datasets employed and the evaluation metrics for assessing our methods. Figure 5 illustrates examples of different datasets and their corresponding ground truth sentences, showcasing the diversity and richness of the data sources utilized in our study. The Microsoft COCO Dataset, with its vast image collection and human-generated captions, serves as a primary resource, complemented by Flickr8k, Flickr30k, and VizWiz datasets, each offering unique attributes for evaluation. We employ BLEU 1-4, Meteor, Rouge-L, and CIDEr metrics to quantify the fidelity of generated captions specifically designed for image captioning tasks. Integrated with MSCOCO tools, these metrics facilitate a comprehensive quantitative analysis, enabling rigorous comparison with existing methodologies and driving advancements in the field of image captioning research

1) MS COCO: The Microsoft COCO Dataset is a massive image recognition, segmentation, and captioning dataset. MS COCO dataset has several features, including object segmentation, recognition in context, multiple objects per class, over 300,000 images, over 2 million instances, and 80 object categories. Each image is accompanied by 5 human-generated captions for that particular image. The dataset is used in the experiments for image captioning methods [5]–[8], [12], [13], [19], [22].

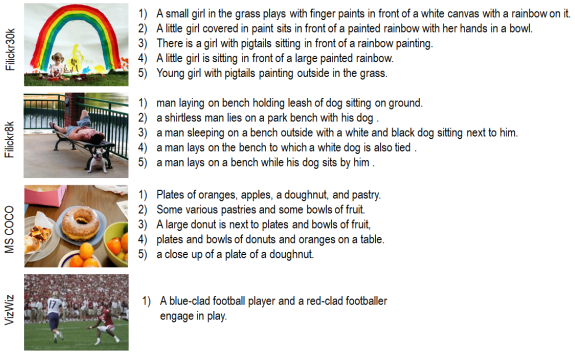

Fig. 5. Examples of image datasets and corresponding captions.

2) Flickr8k: Flickr8k is a popular dataset that contains 8,000 images from Flickr. The training data comprises 6,000 images, while the test and development data comprise 1,000 images each. Humans have annotated five reference captions for each image in the dataset. The dataset contains five reference descriptions for each image, and Table I summarizes the number of images in each dataset. The dataset uses various syntaxes to define the same image, allowing for multiple independent descriptions.

3) Flickr30k: Flickr30k is another popular dataset for image captioning tasks with 31,783 images. For this paper, we are considering splitting 1,000 images for testing, 1,000 for validation, and 29,783 for training the models. Five captions also accompany each image in this dataset.

4) VizWiz: VizWiz-Captions consists of 39,181 images originating from people who are blind that are each paired with 5 captions. VizWiz consists of visual questions asked by blind people seeking answers to their daily visual questions.) Blind photographers capture images, and so are often of poor quality.

TABLE I
DATASETS FOR IMAGE CAPTIONING.

Dataset	Train	Validation	Test	Captions/ Image	Avg.Caption length
MS COCO	82783	40504	40775	5	10.5
Flickr8K	6000	1000	1000	5	11.8
Flickr30K	29,783	1,000	1,000	5	12.3
VizWiz	23,431	7,750	8,000	1	12.6

B. Evaluation Metrics for Image Description

Image captioning metrics were primarily developed to evaluate the text produced by machine translation models and compare the generated captions' quality to the ground truth, as shown in Figure 6. Each metric uses its computation method and has its own set of advantages. The caption evaluation metrics are designed specifically for the image captioning task and evaluate captions generated by captioning image models. Four commonly utilized assessment criteria, namely BLEU 1-4 (Bilingual Evaluation Understudy), Meteor (Metric for Evaluation of Translation with Explicit ORdering), Rouge-L (Recall-Oriented Understudy for Gisting Evaluation -

Longest Common Subsequence), and CIDEr (Consensus-based Image Description Evaluation), are employed to assess the quality of produced sentences using the publicly available MSCOCO tools to analyze the effectiveness of our proposed approaches quantitatively. All n-grams in the generated and reference sentences are compared using these metrics to assess their consistency. To accurately compare with current image captioning methods.

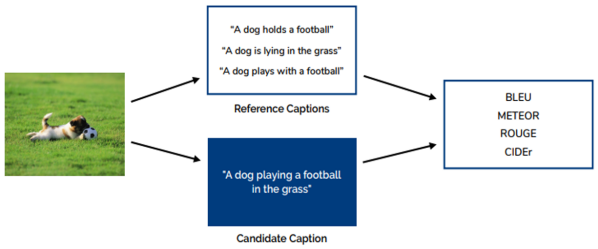

Fig. 6. Evaluation metrics for image description. Assessing caption quality in image captioning models against ground truth references.

VI. Results and Comparison

This section presents quantitative results on the quality of the outputs produced by several image captioning models.

A. Quantitative Analysis

The quantitative results reported in this section demonstrate the efficacy of the proposed method. In Table II to Table V , state-of-the-art models are compared with the proposed method on four datasets: MS COCO, Flickr8k, Flickr30k, and VizWiz. Note that our method outperforms these established benchmarks in multi-comparison.

TABLE II
COMPARISON OF THE PROPOSED MODEL WITH VARIOUS BASELINE APPROACHES ON THE MSCOCO DATASET

MODEL	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
phi-LSTM [8]	0.69	0.52	0.38	0.28	0.51	0.90	0.24
Stim-dri [35]	0.74	0.52	0.36	0.23	0.50	1.04	0.23
r-GRU [34]	0.76	0.57	0.43	0.32	0.54	1.10	0.27
Our	**0.79**	**0.59**	**0.46**	**0.37**	**0.61**	**1.10**	**0.29**

Tables II -V show our model's quantitative comparison and analysis against different baseline approaches. Table II shows that our model performs competitively against phi-LSTMcite [8], Stimulus concept-drivencite [35], and r-GRU (PoS) [34]. Table II shows that our model performs competitively against Self-Critical, phi-LSTM [8], and Top-Down Attention [38]. Table III shows that our model performs competitively against BabyTalk [39], Compositional [37], and attention-seq-to-seq [27].

In Table V, the proposed framework performs more competitively than ACR-RNN [36]. From Tables II-V, our model performs competitively with the current state-of-the-art baseline approaches, which means that our model can increase the accuracy of image captions.

TABLE III
COMPARISON OF THE PROPOSED MODEL WITH VARIOUS BASELINE APPROACHES ON THE FLICKR30K DATASET

MODEL	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
Self-Cri [20]	0.65	0.44	0.30	0.20	0.43	0.43	0.18
phi-LSTM [8]	0.64	0.45	0.31	0.21	0.44	0.45	0.19
Top-Down Att [38]	0.69	0.52	0.38	0.28	0.48	0.57	0.21
Our	**0.71**	0.48	**0.38**	0.27	0.47	0.49	**0.29**

TABLE IV
COMPARISON OF THE PROPOSED MODEL WITH VARIOUS BASELINE APPROACHES ON THE FLICKR8K DATASET

MODEL	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
BabyTalk [39]	0.356	0.218	0.152	0.107	0.261	0.375	0.174
Comp arch [37]	0.639	0.459	0.319	0.217	0.470	.538	0.204
Att Seq2Seq [23]	0.68	0.49	0.41	0.19	0.53	0.41	0.23
Our	**0.69**	**0.51**	0.42	0.21	0.46	0.43	**0.24**

TABLE V
EXPERIMENTAL RESULTS OF STATE-OF-THE-ART METHODS ON VIZWIZ DATASET

MODEL	B-1	B-2	B-3	B-4	Rouge	CIDER	METEOR
ACR-RNN [36]	0.57	0.37	0.24	0.14	0.40	0.30	0.15
our	0.61	0.43	0.29	0.19	0.43	0.27	0.13

The results show that our model can beat the current state-of-the-art baseline approaches and achieve competitive performance against them. Additionally, our model can be used for the image caption task to generate a high-quality image caption compared to the baseline approaches for four image datasets. Figure 7 (a-d) compares the performance of our approach and baseline models in terms of MSCOCO, Flickr30k, Flickr8k, and VizWiz datasets for the image captioning task. The statistical evaluation matrices depicted in the figures indicate the uniform performance of our model on different datasets. This demonstrates that our approach achieves the best results on the image captioning task on different datasets.

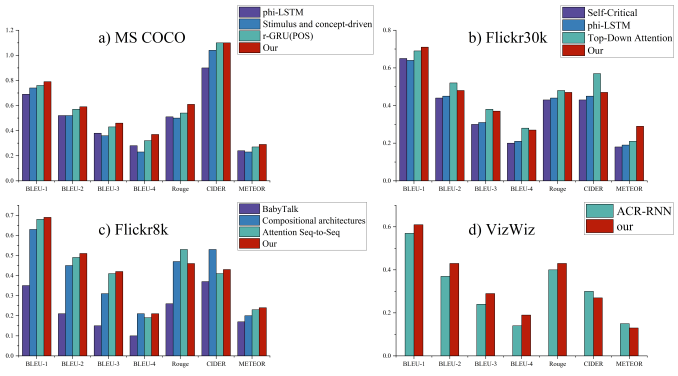

Fig. 7. Evaluation metrics for image description. Assessing caption

B. Qualitative Analysis

of the proposed visual Insight deep multilayer fusion with Inception-based LSTM for descriptive image captioning reveals the promising nature of the model based on multiple datasets. Figure 8 illustrates its application with the MSCOCO dataset,

showcasing its effectiveness in handling diverse visual data. Figure 9 extends this analysis by demonstrating the approach's performance with the Flickr30K dataset, emphasizing its adaptability across different image collections. Furthermore, Figure 10 presents insights gained from employing the approach with the Flickr8K dataset, highlighting its scalability and robustness. Finally, Fig. 11 offers a visualization of the approach's application with the VizWiz dataset

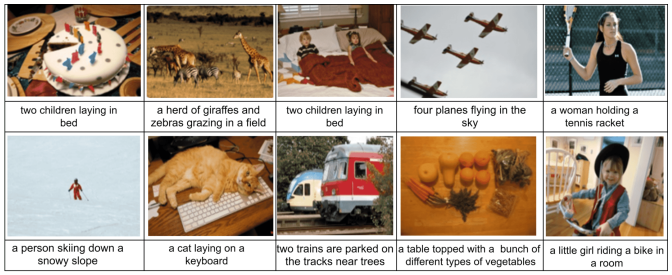

Fig. 8. Visual representation of the proposed approach using the MSCOCO dataset

Although most of the generated captions are informative and contextually accurate, there are instances when the generated caption deviates from the topic of the captioned image. As pointed out above, the errors in depicting the generated caption and the captioned image emanate from two main processes: image recognition and text generation.

Fig. 9. Visual representation of proposed approach using Flickr30k Dataset.

Nonetheless, the proposed model offers numerous benefits. Specifically, the deep multilayer fusion architecture embedded with Inception-based LSTM allows the current model to capture intricate visual features and linguistic intricacies, which ensures that all the generated captions are descriptive and contextual.

Fig. 10. Visual representation of proposed approach using Flickr8k Dataset.

In addition, the proposition of the model's application on multiple datasets indicates the robust nature of the model on numerous image-captioning tasks. The qualitative outcomes, therefore, portray the potential application of the model in image understanding and contextual description.

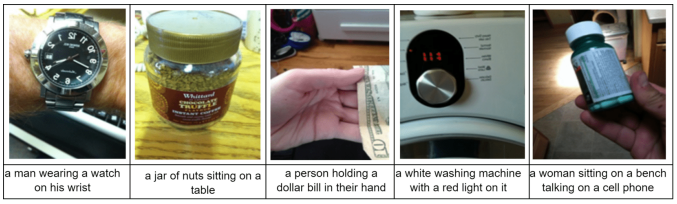

Fig. 11. Visual representation of proposed approach using VizWiz Dataset.

VII. Conclusions

This research study investigates developing an image caption generator using deep learning techniques and integrating CNNs with a recurrent neural network. Applying transfer learning along with pre-existing models such as InceptionV3 and the integration of LSTM architecture has enabled the development of almost correctly matching and topical models to generate captions for the provided images. Finally, the importance of preprocessing, the significance of hyperparameter selection, and the value of evaluation through BLEU1-4, Meteor, CIDEr, and ROUGE. The potential of deep learning demonstrates that it can be used to solve the hard problem of image captioning. While recognizing the limitations and work, mainly in achieving accuracy and diversity, future research areas to be explored include using larger datasets, more sophisticated model architectures, and multimodal integration and connectivity to have the generated models perform better and understand their context better. Ultimately, the work in this research is towards developing state-of-the-art Image captioning technology, with the findings able to foster more innovations across various domains, from computerized vision to accessibility tools, developing technology that can understand and explain images by itself with more human-like fluency and accuracy.

Acknowledgment:
This study was supported by the Project of the Educational Commission of Guangdong Province of China (No. 2022ZDJS113).

References

[1] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156-3164.
[2] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, International conference on machine learning, PMLR, 2015, pp. 2048-2057.
[3] Q. You, H. Jin, Z. Wang, C. Fang, J. Luo, Image captioning with semantic attention, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4651-4659.

[4] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, T. Darrell, Long-term recurrent convolutional networks for visual recognition and description, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2625-2634.

[5] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J.C. Platt, From captions to visual concepts and back, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1473-1482.

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

[7] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 4565-4574.

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

[9] 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.

[10] A. Yuan, X. Li, X. Lu, 3G structure for image caption generation, Neurocomputing, 330 (2019) 17-28.

[11] X. Chen, M. Zhang, Z. Wang, L. Zuo, B. Li, Y. Yang, Leveraging unpaired out-of-domain data for image captioning, Pattern Recognition Letters, 132 (2020) 132-140.

[12] X. He, B. Shi, X. Bai, G.-S. Xia, Z. Zhang, W. Dong, Image caption generation with part of speech guidance, Pattern Recognition Letters, 119 (2019) 229-237.

[13] P. Kinghorn, L. Zhang, L. Shao, A region-based image caption generator with refined descriptions, Neurocomputing, 272 (2018) 416-424.

[14] M. Jamieson, Y. Eskin, A. Fazly, S. Stevenson, S.J. Dickinson, Discovering hierarchical object models from captioned images, Computer Vision and Image Understanding, 116 (2012) 842-853.

[15] C.E. Kahn Jr, D.L. Rubin, Automated semantic indexing of figure captions to improve radiology image retrieval, Journal of the American Medical Informatics Association, 16 (2009) 380-386.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740-755.

[17] D. Gurari, Q. Li, A.J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, J.P. Bigham, Vizwiz grand challenge: Answering visual questions from blind people, Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3608-3617.

[18] T. Ghandi, H. Pourreza, H. Mahyar, Deep learning approaches on image captioning: A review, ACM Computing Surveys, 56 (2023) 1-39.

[19] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A.C. Berg, T.L. Berg, Babytalk: Understanding and generating simple image descriptions, IEEE transactions on pattern analysis and machine intelligence, 35 (2013) 2891-2903.

[20] M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. Berg, T. Berg, H. Daumé III, Midge: Generating image descriptions from computer vision detections, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 747-756.

[21] R. Khan, B. Huang, H. Hassan, A. Zaman, Z. Ye, A Comparative Study of Pre-trained CNNs and GRU-Based Attention for Image Caption Generation, 2023 5th International Conference on Robotics and Computer Vision (ICRCV), IEEE, 2023, pp. 92-99.

[22] X. Chen, C. Lawrence Zitnick, Mind's eye: A recurrent visual representation for image caption generation, Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2422-2431.

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

[24] A. Abdussalam, Z. Ye, A. Hawbani, M. Al-Qatf, R. Khan, NumCap: a number-controlled multi-caption image captioning network, ACM Transactions on Multimedia Computing, Communications and Applications, 19 (2023) 1-24.

[25] H. Wang, H. Wang, K. Xu, Evolutionary recurrent neural network for image captioning, Neurocomputing, 401 (2020) 249-256.

[26] J. Hu, Y. Yang, Y. An, L. Yao, Dual-Spatial Normalized Transformer for image captioning, Engineering Applications of Artificial Intelligence, 123 (2023) 106384.

[27] L.T. Ravulaplli, A Novel Bi-LSTM Based Automatic Image Description Generation, Ingénierie des Systèmes d'Information, 28 (2023).

[28] N. Aafaq, A. Mian, N. Akhtar, W. Liu, M. Shah, Dense video captioning with early linguistic information fusion, IEEE Transactions on Multimedia, 25 (2022) 2309-2322.

[29] H. Wang, H. Wang, K. Xu, Evolutionary recurrent neural network for image captioning, Neurocomputing, 401 (2020) 249-256.

[30] J. Hu, Y. Yang, Y. An, L. Yao, Dual-Spatial Normalized Transformer for image captioning, Engineering Applications of Artificial Intelligence, 123 (2023) 106384.

[31] L.T. Ravulaplli, A Novel Bi-LSTM Based Automatic Image Description Generation, Ingénierie des Systèmes d'Information, 28 (2023).

[32] N. Aafaq, A. Mian, N. Akhtar, W. Liu, M. Shah, Dense video captioning with early linguistic information fusion, IEEE Transactions on Multimedia, 25 (2022) 2309-2322.

[33] X. He, Y. Yang, B. Shi, X. Bai, Vd-san: visual-densely semantic attention network for image caption generation, Neurocomputing, 328 (2019) 48-55.

[34] T. do Carmo Nogueira, C.D.N. Vinhal, G. da Cruz Júnior, M.R.D. Ullmann, Reference-based model using multimodal gated recurrent units for image captioning, Multimedia Tools and Applications, 79 (2020) 30615-30635.

[35] S. Ding, S. Qu, Y. Xi, S. Wan, Stimulus-driven and concept-driven analysis for image caption generation, Neurocomputing, 398 (2020) 520-530.

[36] Ö. Çaylı, V. Kılıç, A. Onan, W. Wang, Auxiliary classifier based residual rnn for image captioning, 2022 30th European Signal Processing Conference (EUSIPCO), IEEE, 2022, pp. 1126-1130.

[37] K. Fu, J. Jin, R. Cui, F. Sha, C. Zhang, Aligning where to see and what to tell: Image captioning with region-based attention and scene-specific contexts, IEEE transactions on pattern analysis and machine intelligence, 39 (2016) 2321-2334.

[38] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and vqa, arXiv preprint arXiv:1707.07998, 2 (2017) 8.

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