ISSN: 2957-9651
Pages: 33-46

Image to Text Description Approach based on Deep Learning Models

Muhanad Hameed Arif*
muhanad.hameed.arif.1981@gmail.com
Directorate of Education in Diyala, Ministry of Education, Diyala, Iraq

Abstract

The image-to-text description can be indicated by creating captions for images that comply with human language perception. Nowadays, with the speedy progress of deep learning models, image-to-text description (or image captioning) has an expanding consideration by numerous researchers in diverse artificial intelligence relevant applications. In general, accurately getting the semantic information of the principal objects in the images and captioning the association among them represents a crucial issue in this field. In this paper, an image-to-text description approach based on Inception-ResNetV2-LSTM with an attention technique is proposed for effective textual descriptions of images.

In this proposed approach, Inception-ResNetV2 is exploited to extract essential features, and the integration of LSTM with the attention technique is implemented as a sentence-creation model in such a way that the learning could be concentrated on specific portions within the images, hence enhancing the performance of image-to-text description approach. In terms of the Meteor and BLEU (1-4) measurements, the proposed approach outperformed other state-of-the-art approaches with 0.787 and (0.977, 0.964, 0.886, and 0.759), respectively.

Keywords: Image-to-Text description, Deep learning approaches, Inception-ResNetV2, LSTM, Attention technique

Article history: Received : 16/12/2023 Accepted : 12/1/2024

1. Introduction

The image-to-text description (or image captioning) represents the comprehension of the images' content and then inferring natural sentences to caption them [1]. The field of image-to-text description works on combining the highly essential prospects in the applications of artificial intelligence (natural language processing and computer vision). This field represented an extremely challenging issue until the emergence of deep learning technology. Individuals are capable of only describing specific images in certain scenarios, hence it has been tricky to attain highly acceptable results. Within the computer vision prospect, in comparison to the task of image classification that appeared relatively mature in the past few years, there are various tricky issues that should be solved in the field of image-to-text description. The process of image description involves several fundamental logical concepts; recognizing objects, specifying ongoing action (activity), and expressing the association among these objects and attributes. All these concepts can be converted into a natural language, and this led to the requirement for the language model to involve visual perception [2].

The issue of image-to-text description is individuated into two fundamental divisions: template-based and retrieval-based approaches (conventional approaches) and deep learning-based approaches. The template-based image-to-text description approaches create captions with pre-defined syntax rules. This category of conventional approaches is not able to create relevant sentences since it is not able to convey visual content accurately. The approaches of retrieval-based image-to-text description retrieve

[*] Corresponding author: muhanad.hameed.arif.1981@gmail.com

the nearest corresponding images and create captions as a description of the inquiry images. This category of conventional approaches utilizes re-ranking to yield the right sentences, however, it fails to alter captions for new images [3, 4].

However, in recent years, the image-to-text description field has been broadly prevalent because of the emergence of deep-learning models in which the structures of the encoder-decoder are utilized for comprehending the images. Fig. 1 demonstrates a typical deep-learning-based approach for image-to-text description [5].Concerning the encoder-decoder image captioning approaches, Convolutional Neural Networks (CNNs) have been exploited as encoders for visual feature extraction from the images, and Recurrent Neural Networks (RNNs), "especially LSTM (Long Short-Term Memory) networks" have been exploited as decoders for transforming the obtained features into various natural languages [6, 7].

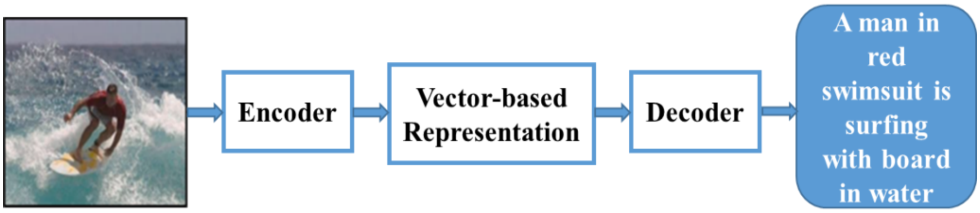

Encoder
Vector-based Representation
Decoder
A man in red swimsuit is surfing with board in water

Fig. 1 Deep learning-based approach for image-to-text description

However, encoder-decoder-based approaches are not capable of analyzing the images over time and considering the spatial prospects of images that are pertinent to the image description (alternatively, creating descriptions for the entire scene). Recently, to conquer the previous restrictions, attention mechanisms have been utilized for mapping text captioning to various areas of the image [8]. In this paper, the principal contributions are specified as follows:

1. An effective image-to-text description approach based on Inception-ResNetV2-LSTM is proposed for recognizing and successfully generating descriptions for the images' objects in the English language.
2. The mechanism of attention is also utilized to enhance the performance of the proposed approach.
3. Comprehensive experiments are examined via utilizing a public dataset and competitive image-to-text description approaches.

The paper's remainder is ordered as follows; the second section outlines the close works related to this paper. The third section explains the proposed approach in detail. The fourth section reports the experiments, attained results, and discussion. The last section demonstrates the fundamental conclusions and future direction.

2. Related Works

Permanently, the image-to-text description field was a significant direction for lots of research, and the approaches for applying such a task can be categorized into; retrieval-based approach, template-based approach, and deep learning-based approach. The first category depends on a broad training set, and the created description is restricted to the description of the training corpus. The second one is restricted to the templates designed manually using a single form without assortment. The final category also depends on a broad training set; however, the created description is not restricted to determined language templates.

Deng et al. [9], proposed an adaptation image-to-text description approach with the assistance of a visual sentry. In the process of image encoding, the proposed approach utilized DenseNet to draw out the comprehensive image features. Simultaneously, the sentry gate was set with an adaptation attention technique to specify whether to employ the feature information of the image for text creation. While in the process of decoding, LSTM was implemented as a language creation method for image description tasks to enhance the image description quality. COCO and Flickr30k datasets were employed in the experiments, and the attained outcomes exhibit considerable progress in METEOR and BLEU (1-4) scores were (0.270) and (0.739, 0.570 0.422, and 0.326) for the COCO dataset, respectively, and (0.214) and (0.667, 0.486, 0.321, and 0.224) for Flickr30K dataset, respectively. However, this proposed approach requires more improvements to attained more accurate outcomes.

Chu et al. [10], presented an automatic description approach using ResNet50-LSTM as encoder-decoder structure. The ResNet50 model generates a representation concerning a given image via including it in a fixed-length vector. The LSTM with the technique of soft attention selectively is concentrated on specific portions in the images for predicting the sentences. This presented approach was trained using the Microsoft COCO (MSCOCO) dataset and assessed using diverse metrics. The attained results of METEOR and BLEU metrics signify the efficiency of the generated good descriptors.

Wang et al. [11], proposed an approach to model the association between the interest regions inside the image using a Graph Neural Network (GNN), in addition to an attention technique for learning relationship-aware visual representation concerning image-to-text description and considering the information of historic context on former attention. The experiments were accomplished using Flickr30K and MSCOCO datasets, and the outcomes illustrated that the proposed approach achieved METEOR of 0.215, and BLEU(1-4) of (0.698, 0.517, 0.378, and 0.277) for the Flickr30K dataset, and METEOR of 0.278, and BLEU(1-4) of (0.759, 0.603, 0.465, 0.358) for MSCOCO dataset

Chang et al. [12], proposed a strengthened approach to image description which involves several fundamental stages such as detecting objects, analyzing colors, and describing images. The first and second stages were accomplished using Mask R-CNN. In the third stage, a pre-trained CNN (called VGG16) was utilized as the encoder, and attention-based LSTM as the decoder. The consolidation of these stages is then implemented to supply more appropriate descriptions for images. Additionally, the created sentences were transformed into speech. This approach was implemented using MSCOCO dataset, and unfortunately, the metrics of assessments were not utilized. To improve the outcomes of image-to-text description, exploiting a generative adversarial network (GAN) is suggested. This GAN can be employed to complete the background of the isolated object image, thereby improving the accuracy of the images' descriptions.

Javanmardi et al. [13], developed an encoder-decoder-based image-to-text description approach using InceptionV3 for extracting features and RNN for creating descriptions. This approach was trained dependent on the MSCOCO dataset and assessed using various metrics. The experiments on this dataset demonstrated that the proposed approach exceeds the other approaches with METEOR of 0.45, and BLEU(1-4) of (0.89, 0.74, 0.61, and 0.54).

Al-Malki and Al-Aama [14], proposed a deep learning-based approach for describing clothing images in Arabic text. A new dataset named Arabic Fashion Dataset (AFD) was created for this context. The first stage of this approach was a multi-label classification of clothing images' attributes. Then Residual Neural Network-50 (ResNet-50) with attention was utilized as an encoder. After that, the LSTM was utilized for learning the words' representation. The last stage (decoder) included a dense 128 layer with the ReLU activation functions for transforming the data accepted from the encoder into a textual sequence.

This approach attained BLEU (1-4) scores of (0.885, 0.838, 0.811, and 0.745). These attained results were promising and indicated that by using larger datasets, the attributes-based approach could provide outstanding findings for Arabic image-to-text description.

Zaidan and Waleed [15] presented an encoder-decoder-based image-to-text description roach using InceptionV3 and LSTM with attention technique. This approach was implemented using the MSCOCO dataset, and the attained BLEU (1-4) scores were (0.543, 0.87, 0.66, and 0.51), and the attained Meteor score was 0.42. However, it was preferable to employ a more refined attention technique like self-attention to refine the model's capability for capturing connections of long-terms between features of an image and generated text.

3.Proposed Approach

The proposed image-to-text description approach encompasses several fundamental phases; inputting the dataset, pre-processing and splitting the dataset, applying a pre-trained model of CNN (Inception-ResNetV2) for extracting features as an encoder, and employing LSTM with attention technique as a decoder for text (sentence) generation. Fig. 3 illustrates the general diagram for the proposed approach

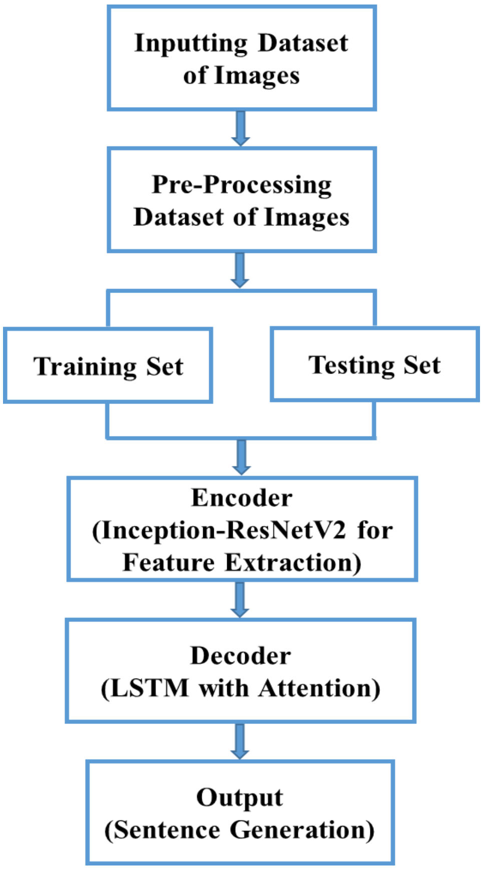
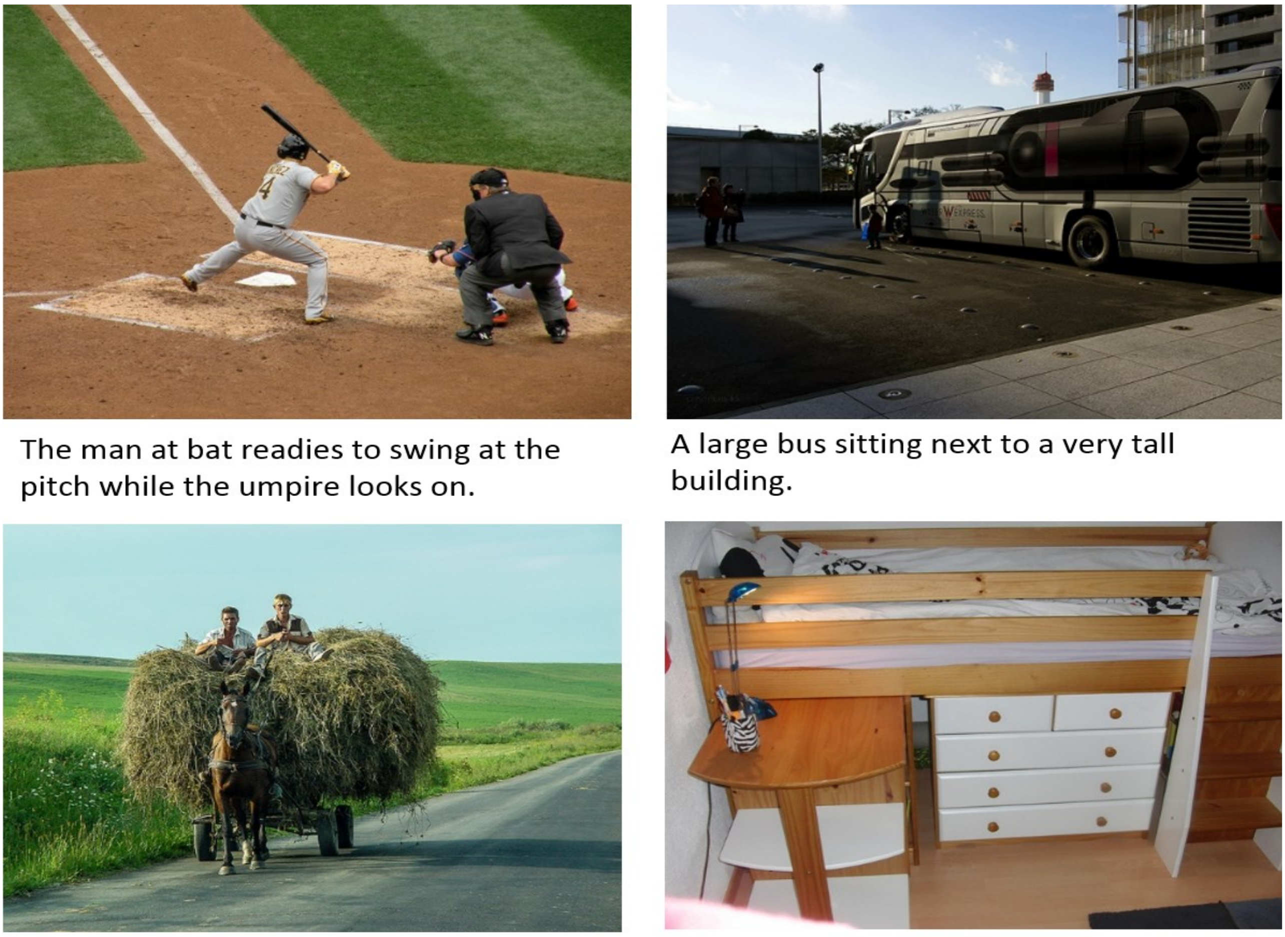

Inputting Dataset of Images
Pre-Processing Dataset of Images
Training Set
Testing Set
Encoder (Inception-ResNetV2 for Feature Extraction)
Decoder (LSTM with Attention)
Output (Sentence Generation)

Fig. 2 General diagram for the proposed approach

A. Dataset Inputting and Pre-processing

In this work, the MSCOCO dataset [16] is utilized. This benchmark dataset was evolved via the team of Microsoft to comprehend scenes, capture images from complicated scenes, and implement multi-tasks like image segmentation, recognition, and description. It utilizes the service of Amazon (Mechanical Turk) to create minimally five sentences for every image (it includes more than "1.5" million sentences). The MSCOCO dataset separated into training, testing, and validation sets that include 154062, 5000, and 5000 images, respectively. Some image samples and descriptions of this dataset are depicted in Fig. 3.



The man at bat readies to swing at the pitch while the umpire looks on.

A large bus sitting next to a very tall building.

A horse carrying a large load of hay and two people sitting on it.

Bunk bed with a narrow shelf sitting underneath it.

Fig.3 Some image samples and descriptions from the MS COCO dataset

After inputting the dataset, the images with varied sizes will be resized into "299×299×3". Then the dataset is split into sets of "80% training" and "20% testing".

B. Phase of Encoding

The encoder (extractor of features) is applied by utilizing Inception-ResNetV2. Inception-ResNetV2 denotes a more resource-intensive hybrid version of Inception, providing significantly improved recognition performance. Unlike the approach of filter concatenation employed in Inception (V3 and V4) architectures, the Inception-ResNetV2 leverage the concept of residual connections. These connections play a crucial role in training deeper network structures, thereby enhancing the whole performance.

The Inception-ResNetV2 structure along with its respective stem, is depicted in Fig. 4. In this diagram, 'V' letter is utilized to signify the utilization of "Valid" padding. Additionally, each layer is accompanied by a numeric value indicating the size of the output. Further insights inside internal modules (Inception blocks) for this version, are depicted in Fig. 5.

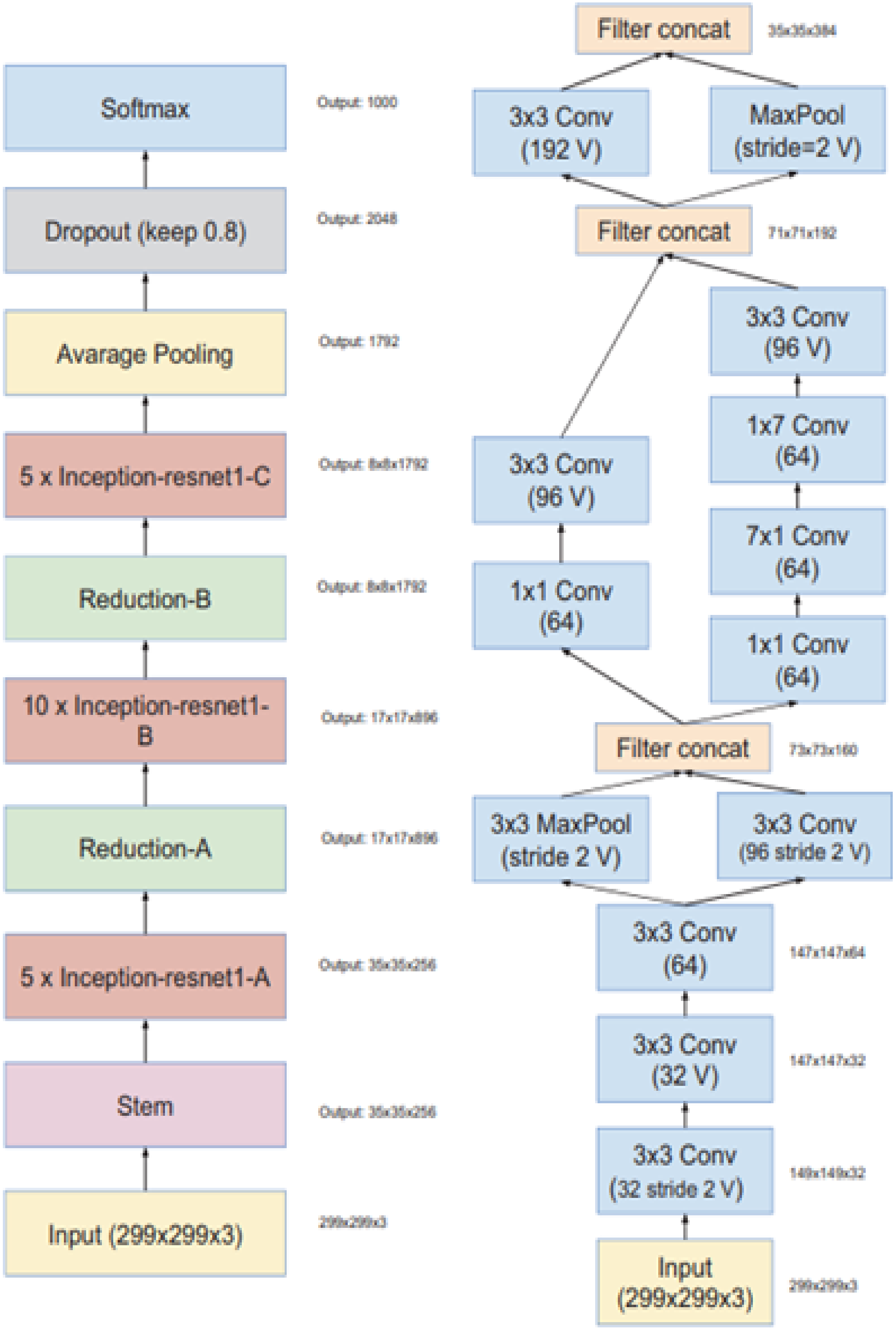
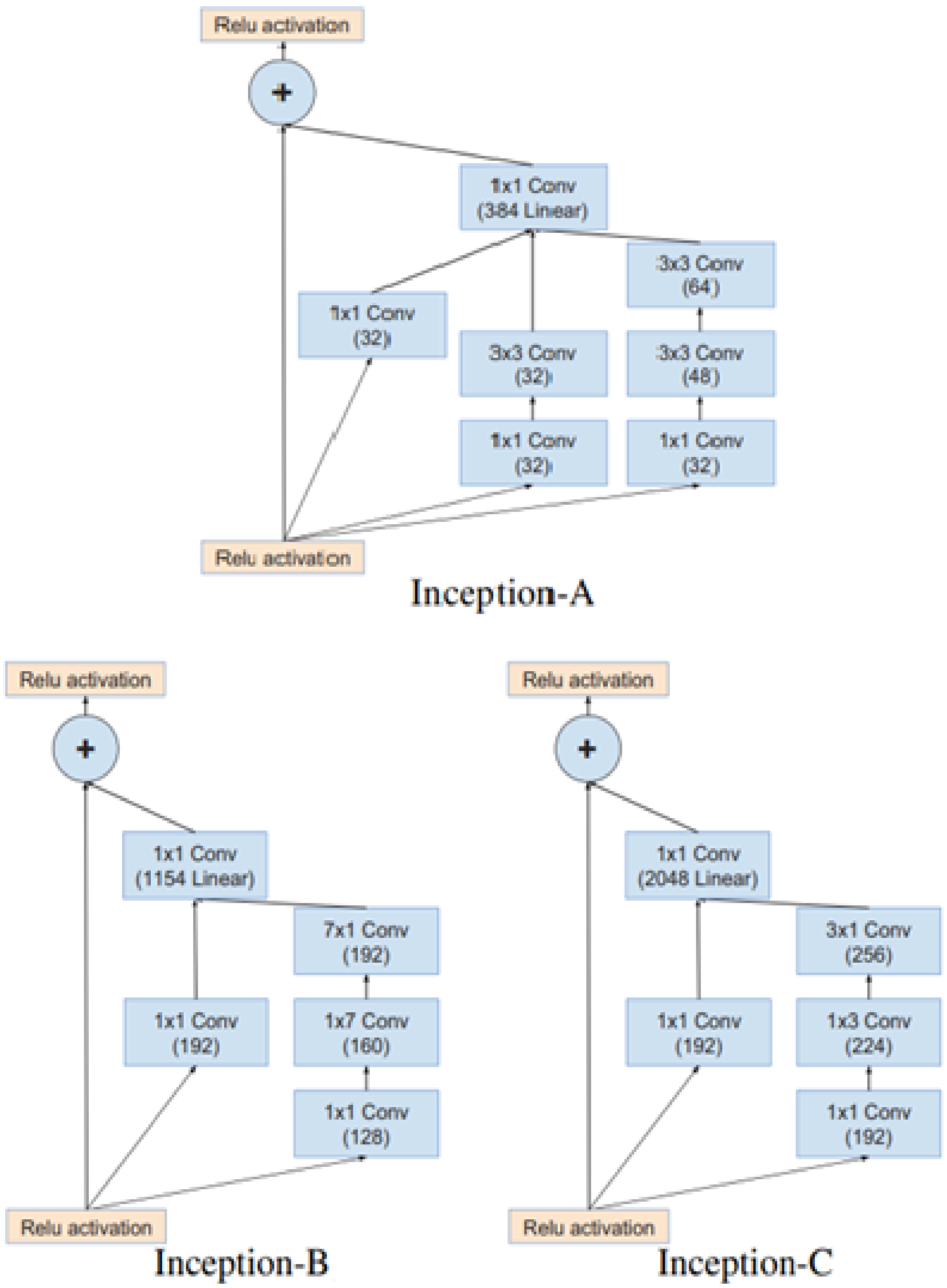


Fig. 4 Inception-ResNetsV2; (a) structure and (b) Its Stem



Inception-A

Inception-B

Inception-C

Fig. 5 Modules within Inception-ResNetV2

C. Phase of Decoding with Attention Technique

The decoding phase utilizes the vector of input to create a sentence that captions the image. This phase is accomplished using LSTM with the technique of visual attention. LSTM is presented for handling the issue of the vanishing gradient existence in the RNN architecture. It provides a memory cell (that enables preserving its state over time) assisted with units termed gates. The typical architecture of LSTM is demonstrated in Fig. 6.

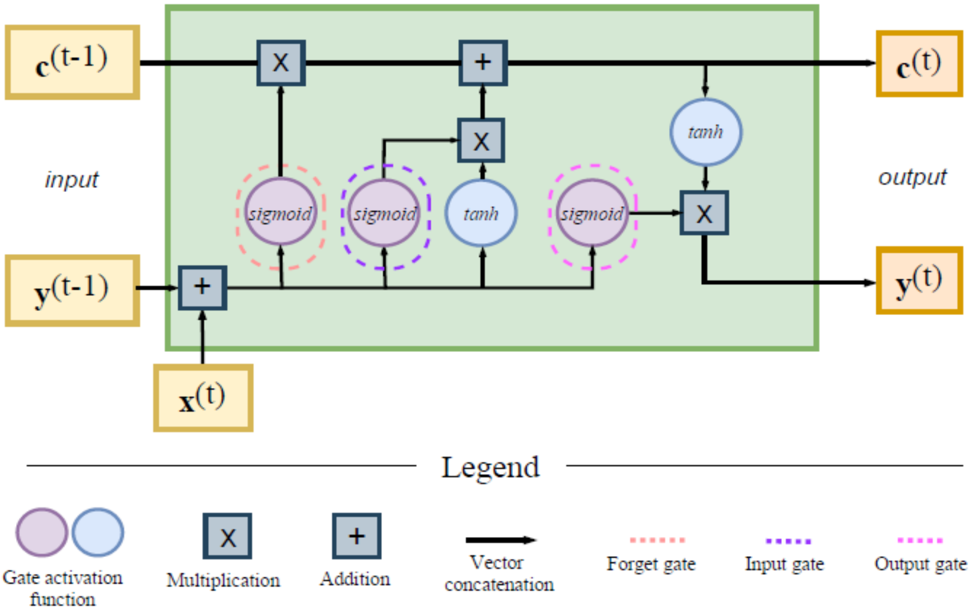


Fig. 6 LSTM architecture

LSTM involves several fundamental components that should be calculated; Firstly, block input works on combining the present input "x$^{(t)}$" and the production (output) "y$^{(t-1)}$" of the LSTM unit in the final repetition; Secondly, input gate works on combining the present input "x$^{(t)}$", the production "y$^{(t-1)}$" of the LSTM unit, and "c$^{(t-1)}$" in the final repetition; Thirdly, forget gate, by which the unit of LSTM can specify the information that must be abandoned from the former cell states "c$^{(t-1)}$". Fourthly, the cell works on combining the input block, and the values of input and forget gates with the former value of the cell. Fifthly, the production gate works on combining the present input "x$^{(t)}$", the production "y$^{(t-1)}$" of the LSTM unit, and the value of the cell "c$^{(t-1)}$" in the final repetition; Finally, the production (output) block works on combining the present value of the cell "c$^{(t)}$" with the value of the present production gate.

The fundamental defect of the approaches that are based on encoder-decoder architecture is that they encode the whole sequence of the input into one fixed-length vector. This is a concern when working with longer sequences of input, especially those that exceed the ones in the set of training. To find the best solution to this issue, the technique of visual attention is utilized which contrary to the conventional encoder-decoder architectures, works on developing a special vector for every decoder time pace. In this proposed approach, the attention technique developed in [17] is utilized.

This technique is based on input and output components and allows efficient employment of the decoder of the highest related information for the sequence of the input. This is accomplished by utilizing the whole outputs of the encoder to construct a weighted set wherever the maximum weights are designated to the highest related vectors. All the phases of the proposed approach are demonstrated in detail in Fig. 7.

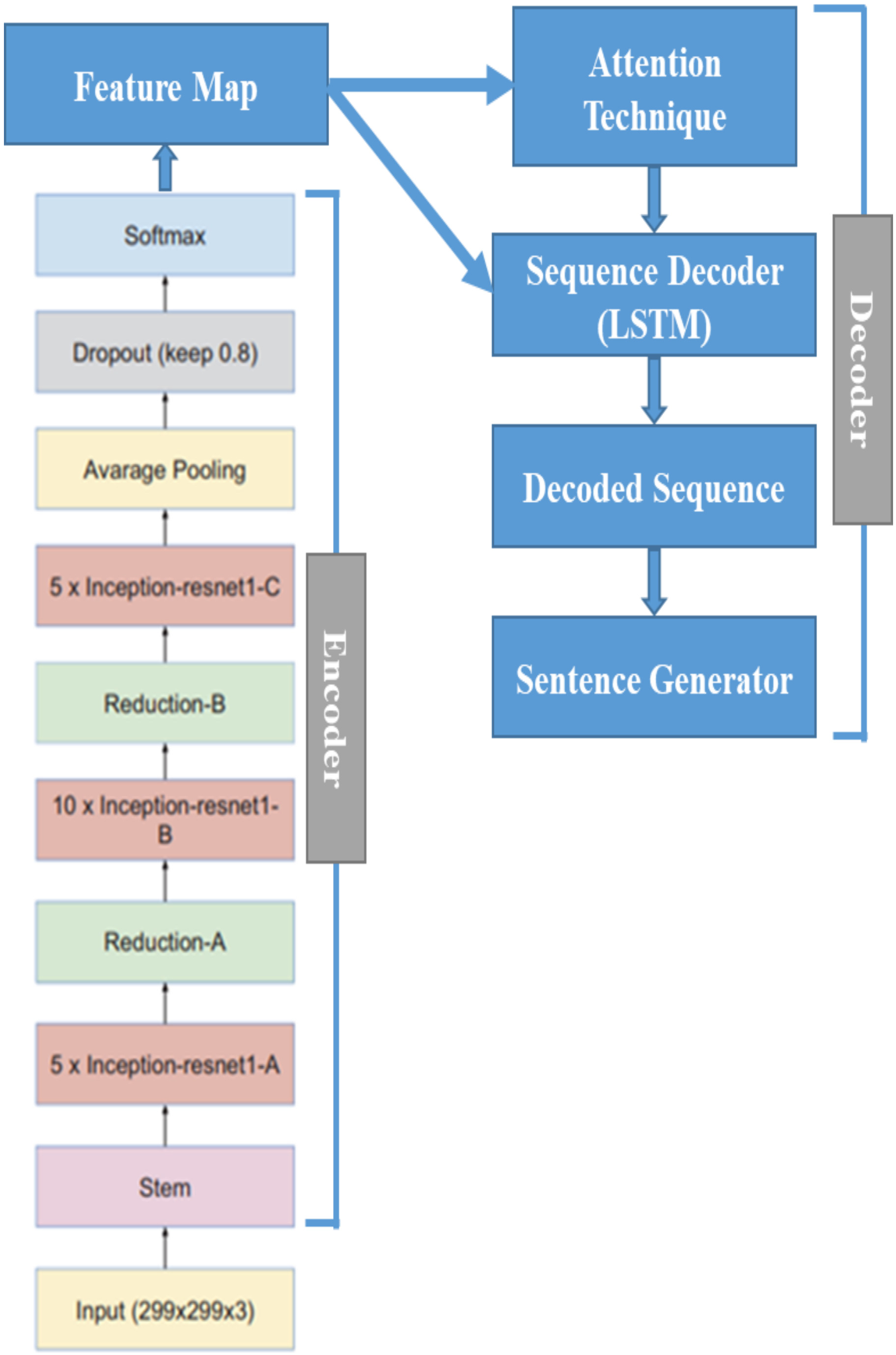


Fig. 7 The architecture of the proposed approach using Inception-ResNetV2-LSTM with an attention technique

4.Results and Discussion

The utilized dataset for text-to-image description involves a collection of images paired with descriptions in the English language (typically scripted by people) that explain their content. Many metrics utilized for assessing text-to-image description approaches were dependent on these reference sentences. Consequently, considering a sentence created by the proposed approach, the metrics assess the created sentence by calculating its likeness versus a collection of reference sentences related to a certain visual content. In this work, BLEU and METEOR are utilized as reference-based metrics for assessing the performance of the proposed approach.

BLEU [18] indicates a low-cost, fast, and language-unrelated metric originally presented for auto-assessment of machine translation, and it is typically utilized for assessing image description approaches. This metric calculates the m-grams overlapping precision for the created sentence with the reference descriptions, and it is generally utilized with m-grams of size (1-4). The scores of BLEUS range from zero to one and scores over 0.30 are usually considered an intelligible sentence, and over 0.50 are considered acceptable candidate sentences. This metric is calculated as follows:

$$S_p = \begin{cases} e^{(1-l_r/l_c)} & if\ l_c \leq l_r \\ 1 & if\ l_c > l_r \end{cases} \quad (1)$$

$$BLEU = S_p \cdot exp\ (\textstyle\sum_{m=1}^{M} W_m \log G_m) \quad (2)$$

Where S_p indicates the factor of a shortness penalty for penalizing created sentences that are shorter than reference sentences, l_c indicates the created sentence length, l_r indicates the reference collection length, exp indicates the function of exponential, W_m indicates positive weights that sum to one (usually set to $1/M$), and G_m indicates the geometric averaging of the adjusted m-gram precisions till M (M is set to 4). The attained scores of BLEU (1-4) under various splitting percentages (of training and testing) for the dataset are demonstrated in Fig. 8.

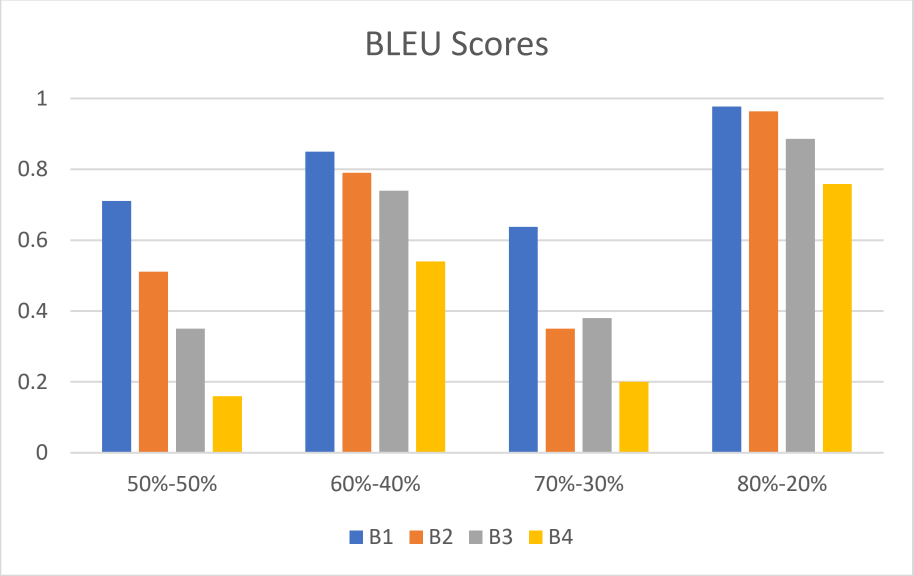


Fig. 8 Scores of BLEU under various splitting percentages

The METEOR metric [19] was presented to handle the flaw observed in the above-mentioned metric, involving the utilization of m-grams with higher-order, the utilization of G_m for m-grams, the absence of recall, and the absence of clear word-matching between created and reference sentences. METEOR is dependent on the harmonic mean, recall, and precision, and it involves generating alignment between uni-grams from created and reference sentences. This metric is calculated as follows:

$$METEOR = \frac{10RePr}{9Pr+Re}(1 - Pen) \quad (3)$$

$$Pen = 0.5(\frac{C_c}{C_u})^3 \quad (4)$$

Where Re and Pr symbolize Recall and Precision, and these measures are calculated as n/Ur and n/Uc, respectively. n indicates the count of uni-grams occurring in created and reference sentences, and Ur and Uc indicate the count of uni-grams in the reference sentence and the count of uni-grams in the created sentence, respectively. Pen indicates the penalty, C_u indicates the total count of matched uni-grams, and C_c indicates the minimum potential count of chunks (collections of matched uni-grams that arise in a similar arrangement in the created and reference sentences. The attained results of Meteor under various splitting percentages (of training and testing) for the dataset are demonstrated in Fig. 9.

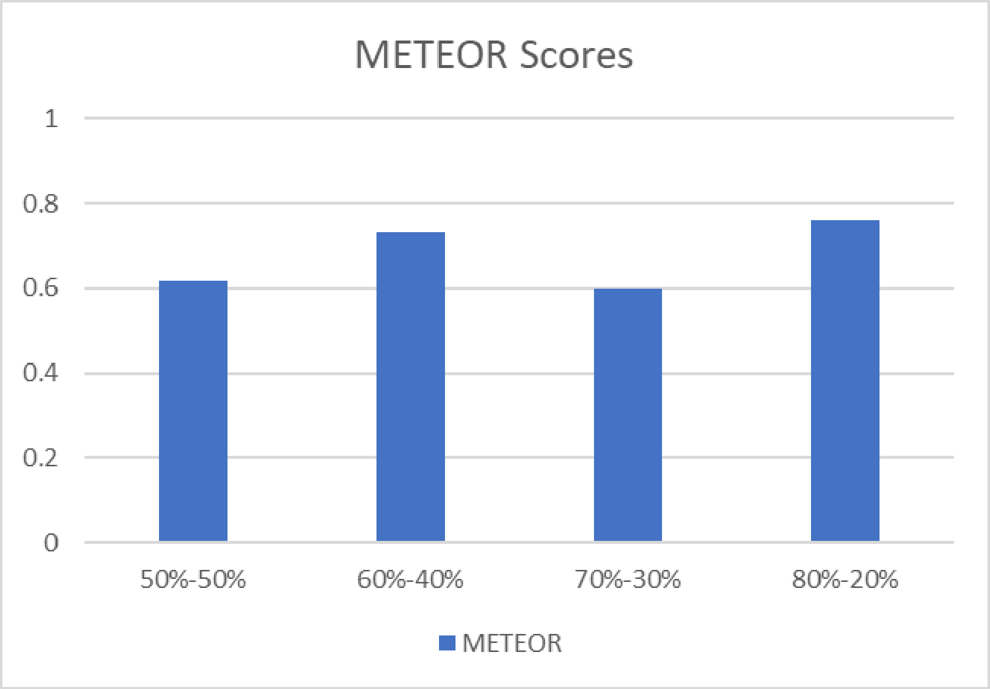


Fig. 9 scores of METEORS under various splitting percentages Curves of losses

Fig. 10 depicts the progression of the (validation and training) losses.

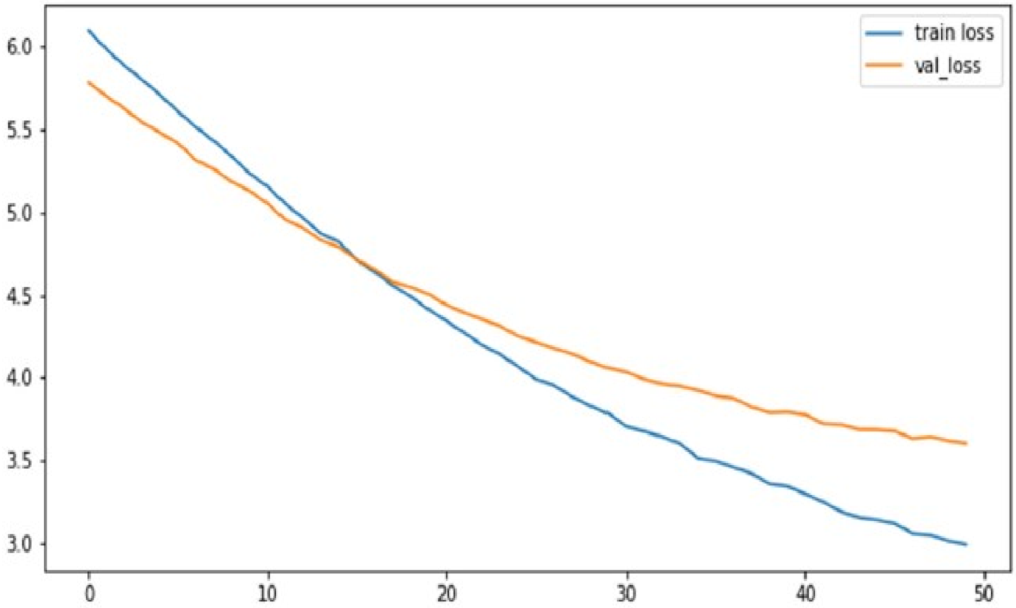


Fig. 10 Curves of losses

Considering the above curves of losses, we can observe that the curve of the training loss is declining faster compared with the other curve of loss (over 50 epochs). The proposed approach outperformed other state-of-the-art approaches that utilized MSCOCO dataset (as demonstrated in Table 1) with 0.787 and (0.977, 0.964, 0.886, and 0.759), for Meteor and BLEU (1-4) scores, respectively

Table 1 Proposed approach versus related approaches

Ref.	Approaches	METEOR Score	BLEU (1-4) Scores
[9]	Dense Net-LSTM with attention	0.270	0.739, 0.570 0.422, and 0.326
[10]	ResNet50-LSTM with soft attention	0.261	0.731, 0.562, 0.41, 0.326
[11]	GNN with context-aware attention	0.278	0.759, 0.603, 0.465, 0.358
[12]	VGG16-LSTM with attention	-	-
[13]	InceptionV3-RNN	0.45	0.89, 0.74, 0.61, and 0.54
[15]	InceptionV3-LSTM with attention	0.42	0.543, 0.87, 0.66, and 0.51
Proposed approach	Inception-ResNetV2-LSTM with an attention	0.787	0.977, 0.964, 0.886, and 0.759

The example for image-to-text description (depicted in Fig. 11) explains how the attention technique can create accurate descriptions by enabling the approach to concentrate on the highest related area within the image at every time step and generate a description. These created descriptions involved further clarification concerning the objects and their relations, and additionally, defined the semantic connections between the goal object and the scene in the image.

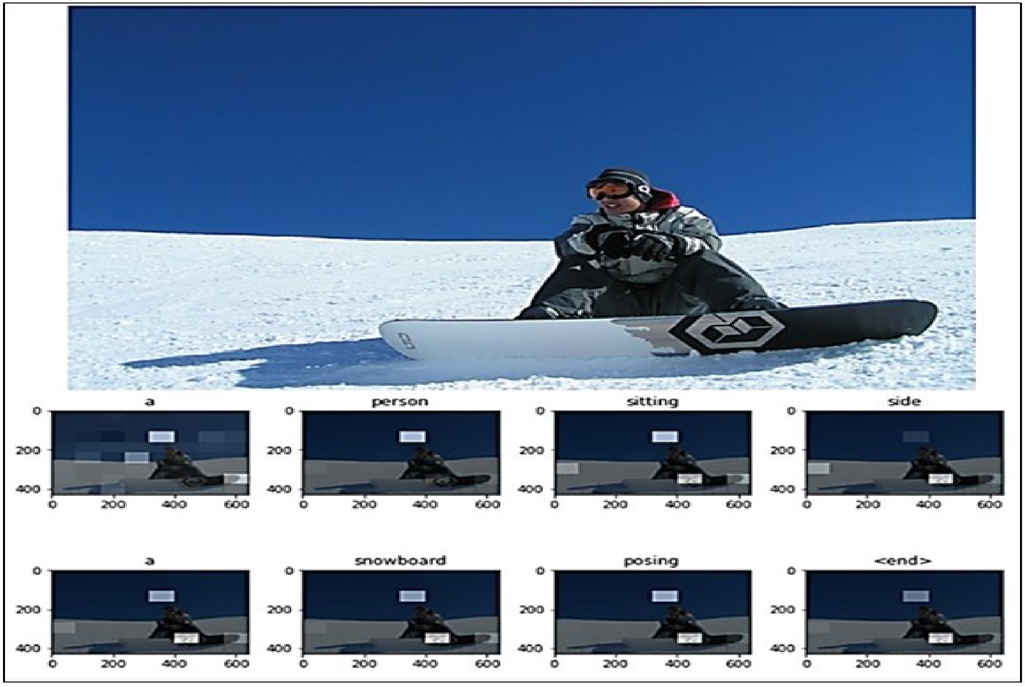


Fig. 11 An example of the image-to-text description

5. Conclusion

In this work, an effective image-to-text description approach based on encoder–decoder with attention technique is proposed. The Inception-ResNetV2 network was utilized in this proposed approach as an alternative to the conventional models of CNNs for improving the global feature extraction of images and improving text description accuracy. These more comprehensive features were then transmitted to the decoder (LSTM with an adjusted attention technique) for generating textual sentences. This proposed approach reached satisfactory enough outcomes in BLEU and METEOR evaluation terms and has attained.

The fundamental aim of enhancing the accuracy and generalization of the approach. In the future, we will focus on generating real-time image-to-text descriptions. And it is possible to apply the proposed approach to other tasks like answering the visual question. Furthermore, the generation of captioning for 3-D images represents an encouraging research direction in the future.

References

[1]. Runyan, D., Wenkai, Z., Zhi, G., Xian, S. (2023). A Survey on Learning Objects' Relationship for Image Captioning", Computational Intelligence and Neuroscience, 2023, 1-16.

[2]. Cheng, C., Li, C., Han, Y., Zhu, Y. (2021). A semi-supervised deep learning image caption model based on Pseudo Label and N-gram", International Journal of Approximate Reasoning, 131, 93-107.

[3]. Oluwasammi, A., Aftab, M. U., Qin, Z., Ngo, S. T., Doan, T. V., Nguyen, S. B., Nguyen, S. H., Nguyen, G. H. (2021). Features to Text: A Comprehensive Survey of Deep Learning on Semantic Segmentation and Image Captioning", Complexity, 1-19.

[4]. Jia, J., Ding, X., Pang, S., Gao, X., Xin, X., Hu, R., Nie, J. (2023). Image captioning based on scene graphs: A survey", Expert Systems with Applications, 231.

[5]. Sharma, D., Dhiman, C., Kumar, D. (2023). Evolution of visual data captioning Methods, Datasets, and evaluation Metrics: A comprehensive survey. Expert Systems with Applications, 221.

[6]. Singh, A., Raguru, J. K., Prasad, G., Chauhan, S., Tiwari, P. K., Zaguia, A., Ullah, M. A. (2022). Medical Image Captioning Using Optimized Deep Learning Model", Computational Intelligence and Neuroscience, 2022, 1-9.

[7]. Albawi, S., Arif, M. H., & Waleed, J. (2022). Skin cancer classification dermatologist-level based on deep learning model. Acta Scientiarum. Technology, 45(1), e61531.

[8]. Poddar, A. K. & Rani, R. (2023). Hybrid Architecture using CNN and LSTM for Image Captioning in Hindi Language. Procedia Computer Science, 218, 686-696.

[9]. Deng, Z., Jiang, Z., Lan, R., Huang, W., Luo, X. (2020). Image captioning using DenseNet network and adaptive attention. Signal Processing: Image Communication, 85, 1-9.

[10]. Chu, Y., Yue, X., Yu, L., Sergei, M., Wang, Z. (2020). Automatic Image Captioning Based on ResNet50 and LSTM with Soft Attention. Wireless Communications and Mobile Computing, 2020:1-7.

[11]. Wang, J., Wang, W., Wang, L., Wang, Z., Feng, D. D., Tan, T. (2020). Learning visual relationship and context-aware attention for image captioning. Pattern Recognition. 98: 107075.

[12]. Chang, Y.-H., Chen, Y.-J., Huang, R.-H. , Yu, Y.-T. (2022). Enhanced Image Captioning with Color Recognition Using Deep Learning Methods. Applied Sciences, 12(1): 209.

[13]. Javanmardi, S., Latif, A.M., Sadeghi, M.T., Jahanbanifard, M., Bonsangue, M., Verbeek, F.J. (2022). Caps Captioning: A Modern Image Captioning Approach Based on Improved Capsule Network. Sensors, 22, 1-20.

[14]. Al-Malki, R. S., & Al-Aama, A. Y. (2023). Arabic Captioning for Images of Clothing Using Deep Learning". Sensors, 23(8).

[15]. Zaidan, Y. H., & Waleed, J. (2023). Image Captioning Generation Using Inception V3 and Attention Mechanism. Journal of Al-Qadisiyah for Computer Science and Mathematics, 15(2), 1-10.

[16]. Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollar, p., Zitnick, C. L. (2015). Microsoft COCO Captions: Data Collection and Evaluation Server. Computer Vision and Pattern Recognition. arXiv:1504.00325 [cs.CV].

[17]. Bahdanau, D., Cho, K., Bengio, Y., 2015. Neural machine translation by jointly learning to align and translate. In: 2015, 3rd International Conference on Learning Representations, ICLR 2015.

[18]. Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In Proc. of the 40th annual meeting of the association for computational linguistics, 311-318.

[19]. Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proc. of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/Or summarization, 65-72.