

The Role of Attention Mechanism in Generating Image Captions: An Innovative Approach with Neural Network-Based Seq2seq Model

Zeynep Karaca ¹ , Bihter Daş ² 

¹ Department of Software Engineering, Technology Faculty, Firat University; Elazığ, Türkiye

² Department of Software Engineering, Technology Faculty, Firat University; Elazığ, Türkiye

Corresponding author:

Bihter Das, Firat University
Department of Software Engineering,
Technology Faculty, Elazığ/Turkey
bihterdas@firat.edu.tr

Article History:
Received: 09.08.2023
Accepted: 26.03.2024
Published Online: 26.03.2024

ABSTRACT

This study addresses important contributions to generating text from images, aiming to create meaning in various fields such as entertainment, communication, commerce, security, and education by establishing a connection between visual and textual content. This process aims to increase the accessibility, understandability, and processability of content by converting image data into meaningful text. Therefore, advances and studies in this field are extremely important. This study focuses on the effect of the combination of deep neural network models and attention mechanisms in creating more meaningful captions from images. Experiments performed on the Flickr8k dataset highlight the abilities of Seq2seq and VGG19 models to generate titles compatible with reference sentences. By using the dynamic focusing feature of the attention mechanism, the model effectively captures detailed aspects of images. The findings of this study have the potential to push the boundaries of multimodal data processing and representation with the effective integration of visual and textual information by adding information that the attention mechanism works more effectively together with the Seq2seq model.

Keywords: Image-to-text, Attention mechanism, Seq2seq model, VGG19, Image Capturing, Deep learning

1. Introduction

Natural Language Processing (NLP) has evolved into a significant field within the realm of artificial intelligence, focusing on the automatic analysis and representation of human language through computational algorithms [1]. NLP finds effective applications in various domains, such as understanding, translation, summarization, and sentiment analysis of texts. Particularly in today's rapidly advancing technological landscape, innovations in NLP hold paramount importance [2].

In this context, the creation of image captions stands as a critical domain at the intersection of Natural Language Processing, computer vision, and artificial intelligence [3]. Image captioning involves the ability to narrate image content using language, and research in this field serves essential purposes, such as improving the quality of life for visually impaired individuals and enhancing overall human-computer interaction [4].

With the proliferation of images on the internet and the advancements in artificial intelligence technologies, the significance of generating image captions has surged. Among the fundamental challenges in this domain is not only accurately and meaningfully describing the content of an image but also expressing the relationships between objects present within the image [5].

The process of generating image captions involves a combination of components from deep learning. Convolutional Neural Networks (CNNs) are employed for image processing, encoding images to extract features, while models like Recurrent Neural Networks (RNNs) are used for natural language processing to generate descriptions. This necessitates a successful integration of both visual and linguistic models [6,7].

The manuscript aims to show how much the use of the attention mechanism with a deep learning-based model contributes to the performance of the system. The proposed approach is demonstrated on a data set widely used in the literature because an accurate comparison with existing methods was desired. In the study, the performance of the attention mechanism was shown with two different deep neural network models. The attention mechanism was used with the VGG19 deep learning model in addition to the Seq2seq model, and the contribution of the proposed approach to the performance of the system was tested

with different methods. This study proposes a combined approach with a neural network-based model and attention mechanism for generating image captions. The suggested first neural network-based model employs an encoder-decoder architecture based on the Sequence-to-Sequence (Seq2seq) framework, with the incorporation of an attention mechanism to enhance performance. The second model is the VGG19 model, which is a deep model with 19 layers. The evaluation of the models is carried out using widely adopted metrics such as BLEU, METEOR, and ROUGE. In this study, current developments in image captioning are examined in depth, and the performance of the proposed approach and different methods are discussed.

The rest of the paper is organized as follows. Section 2 provides a brief overview of the existing research conducted in this field. In Section 3, the utilized dataset, ESA model, Seq2Seq model, encoder-decoder architecture, and attention mechanism employed in the experimental methodology are discussed. Section 4 presents the experimental results concerning the proposed method, and finally, the overall contributions of the study are presented in the Conclusion section.

2. Literature Review

Previous studies on image-to-text generation highlight the effective integration of deep learning and natural language processing techniques. This section extensively examines these approaches in the literature. Various studies have explored the generation of image captions, highlighting diverse approaches.

Yue et al. [8] proposed a method for Thangka image recognition based on an attention mechanism and encoder-decoder architecture. They utilized the ResNet101+ Convolutional Block Attention Module (CBAM) to extract image features and incorporated an attention mechanism into the encoder for improved feature extraction. Using Long Short-Term Memory (LSTM) as the decoder, they validated the model using the Thangka dataset and Flickr8k dataset. The results demonstrated the higher success of the ResNet101 + CBAM model compared to ResNet101.

Chandaran et al. [9] suggested a YOLOv5 model combined with Bidirectional Long Short-Term Memory (Bi-LSTM) for object recognition and feature extraction. They evaluated performance using the Bleu metric and achieved favorable outcomes.

Bhadauria et al. [10] combined two different LSTM networks with CNN using the Flickr8k dataset. Due to the small dataset size, accuracy levels were not significantly high.

Shaikh et al. [11] adopted an alternative approach using an encoder-decoder architecture. They employed Convolutional Neural Networks (CNN) as the encoder and Gated Recurrent Unit (GRU) as the decoder. The results indicated that the GRU model with an open neighborhood connection exhibited higher success.

Xue et al. [12] proposed a model combining the cloze-style approach with neural networks. They employed WGAN to generate sentence templates with broader visual coverage and utilized CNN to fill in gaps in visual regions using object detectors. The model was trained on Flickr8k and MSCOCO datasets, showcasing superior performance.

Singh et al. [13] utilized a hybrid model with CNN and LSTM, trained on the Flickr8k dataset, achieving a Bleu score of 0.52.

Han et al. [14] employed an interpretable image captioning generator model based on determining why specific objects are present in an image. Their model featured a generation module and an explanation module, using an encoder-decoder architecture to generate captions. The explanation module constructed a weight matrix for all words in the generated Caption from detected regions in the image.

Wang et al. [15] utilized an end-to-end deep learning model with a semantic attention mechanism for caption generation. They calculated the similarity between end-to-end frame image feature sequences and semantic word sequences using a derived structure. The model was applied by transferring English information from the Flickr8k dataset to Chinese, showcasing a 3.9% improvement over state-of-the-art approaches. In Chinese captions, the model achieved Bleu-1 63.7, Bleu-2 49.4, Bleu-3 37.2, Bleu-4 28.7, Rouge_L 53.34, and CIDEr 51.45, generally higher than English except for Bleu-1. On the MS COCO dataset, the model achieved Bleu-1 73.1, Bleu-2 55.9, Bleu-3 43.4, Bleu-4 32.8, Rouge_L 25.49, and CIDEr 95.1 values.

Chen et al. [16] introduced the SGGC (Scene Graph Guiding Captioning) model, aiming to bridge the semantic gap between scene graphs, images, and captions to generate better sentences. The model employed scene graphs for decoder generation and used an attention mechanism for caption production. It was evaluated on MS COCO and Flickr30k datasets, with MS COCO achieving Bleu-1 77.2, Bleu-2 60.7, Bleu-3 46.2, Bleu-4 36.3, Meteor 27.8, CIDEr 116.5, and Rouge_L 56.7 scores. For the Flickr30k dataset, it obtained Bleu-1 66.9, Bleu-2 49.4, Bleu-3 35.1, Bleu-4 24.8, Meteor 20.3, CIDEr 116.5, and Rouge_L 53.3 scores.

Rafi et al. [17] utilized Linear Substructures, a model understanding word relationships and sequential orders for caption generation. They focused on regions of motion in images using a common variance shift and employed Inceptionv3

architecture with LSTM encoders to extract image features. Flickr8k dataset was used, resulting in improved performance compared to CNN-based approaches, with Bleu-1 at 81.1 and Bleu-2 at 38.6.

Sharma et al. [18] employed a Lightweight Transformer architecture with a GRU-integrated decoder. They reduced the standard encoder-decoder architecture to an Inceptionv3 + Transformer encoder and Transformer decoder. Model evaluation on the MS COCO dataset yielded Bleu-1 81.0, Bleu-2 65.2, Bleu-3 65.2, Bleu-4 37.8, Meteor 27.9, CIDEr 123.1, and Rouge 58.0 scores.

Panigrahi et al. [19] proposed a model leveraging relationships between image regions, caption words, and RNN model states. VGG16 was used as an encoder to convert image regions to feature vectors, which were then used as inputs to an LSTM-based decoder. The LSTM decoder predicted word sequences to generate descriptions. The model, trained for 20 epochs on the Flickr8k dataset, achieved 20.28% loss and 85% accuracy.

Sun et al. [20] proposed the Bidirectional Beam Search (BiBS) method for bidirectional inference. They introduced the method of Gap-filling Image Captioning, considering past and future sentence structures to obtain accurate image captions. The Visual Madlibs dataset was used. Tested on the MS COCO dataset, the method outperformed baseline techniques. The BiRNNBiBS method achieved Bleu-1 0.470 and Bleu-2 0.389 scores.

Keneshloo et al. [21] combined Reinforcement Learning and the seq2seq model to integrate decision-making and long-term memory. They employed their own models for training the DCA (SCPG) based model, which achieved Rouge-1 41.69, Rouge-2 19.47, and Rouge-L 37.92 scores.

Sahrial Alam et al. [22] evaluated performance using five different models (VGG16, ResNet50, InceptionV3, DenseNet201, Xception) on the Flickr8k dataset. Model accuracies were: VGG16 0.83, ResNet50 0.87, InceptionV3 0.80, DenseNet201 0.87, and Xception 0.81.

Zhou et al. [23] proposed the Triplet Sequence Generative Adversarial Networks (TSGAN) model for unsupervised image captioning. Tested on the MSCOCO dataset, it achieved Bleu-1 46.2, Bleu-2 26.8, Bleu-3 13.5, Bleu-4 6.9, Meteor 13.0, Rouge 32.3, CIDEr 28.9, and Spice 8.3 scores.

Kushwaha et al. [24] utilized the VGG19+LSTM model, extracting image features using the VGG19 model. Compared with other models on the Flickr8k dataset, the VGG16+LSTM model received the highest Bleu score of 55.9.

Liu et al. [25] proposed the Vocabulary-Critical Sequence Training (VCST) method, a novel Reinforcement Learning approach assigning distinct values to words at each step. MS COCO 2014 dataset was used, and the VCST method was applied with the SCST training method for ATT2in, Top-Down, Up-Down, and SGAE models. The UpDown+SCST+VCST model achieved Bleu-4 38.0 and CIDEr-D 125.0 scores.

Zheng et al. [26] proposed the Di-vCon method, which generates explanations with various semantic concepts. The method consists of two steps. In the first step, a concept sequence generator is developed to automatically generate concept sequences in reverse order. The second step involves a sentence generator that takes concept sequences as input and produces descriptions for each sequence. The model focuses more on less frequent objects and achieves optimal performance on the MS COCO dataset with beam size two and group size 1.

Table 1 Some of the Studies for Text Production from Images

| Author | Dataset | Method | Results |
|----------------------|-----------|--------------------------|--|
| Singh et al. [13] | FLICKR8K | CNN+LSTM | BLEU=0.52 |
| Chen et al. [16] | MSCOCO | SGGC+Attention Mechanism | BLEU1=77.2 BLEU2=60.7 BLEU3=46.2 BLEU4=36.3 METEOR=27.8 CIDEr=116.5 ROUGE-L=56.7 |
| | FLICKR30K | | BLEU1=66.9 BLEU2=49.4 BLEU3=35.1 BLEU4=24.8 METEOR=20.3 CIDEr=116.5 ROUGE-L=53.3 |
| Rafi et al. [17] | FLICKR8K | InceptionV3+LSTM | BLEU1=81.1 BLEU2=38.3 |
| Sharma et al. [18] | MSCOCO | InceptionV3+Transformer | BLEU1=81.0 BLEU2=65.2 BLEU3=65.2 BLEU4=37.8 METEOR=27.9 CIDEr=123.1 ROUGE-L=58.0 |
| Zhou et al. [23] | MSCOCO | TSGAN | BLEU1=46.2 BLEU2=26.8 BLEU3=13.5 BLEU4=6.9 METEOR=13.0 CIDEr=28.9 ROUGE=32.3 SPICE=8.3 |
| Kushwaha et al. [24] | FLICKR8K | VGG16+LSTM | BLEU=55.9 |

Birmingham et al. [27] introduced the KENGIC model for image captioning based on keyword and n-gram graph. A series of image keyword nodes are connected through overlapping n-grams to form a directed graph, generating titles from the most probable n-gram sequences. MS COCO dataset was used. The KENGIC model's performance results on the COCO Karpathy dataset are Bleu-1 66.3, Bleu-4 18.6, Meteor 22.6, Rouge 40.4, CIDEr 67.8, and Spice 18.5.

Dwivedi et al. [28] used a 5-layer CNN for extracting image features and an RNN for processing text data. The proposed CNN-5 model was compared with transfer learning models VGG-16 and VGG-19 using the MNIST dataset. CNN-5 model achieved Bleu-1 0.5348, Bleu-2 0.2357, Bleu-3 0.1488, and Bleu-4 0.0660 scores. VGG-16 model achieved Bleu-1 0.5362, Bleu-2 0.2566, Bleu-3 0.1432, and Bleu-4 0.0591 scores. VGG-19 model achieved Bleu-1 0.5240, Bleu-2 0.2430, Bleu-3 0.1370, and Bleu-4 0.0511 scores. The highest Bleu-1 and Bleu-2 values were in the VGG-16 model, while the highest Bleu-3 and Bleu-4 values were in the CNN-5 model. Table 1 lists some of the existing works for image-to-text generation.

In the general evaluation of studies in the literature, the diversity observed between different methods and data sets is striking. The change in performance values depending on the methods used in the studies shows that Transformer-based models, such as InceptionV3+Transformer [18], are more successful than other models. Transformer architecture has shown impressive results in the field of generating text from images, especially with its success in natural language processing tasks. This success can be specifically attributed to the ability to better capture meaning in language. On the other hand, the data sets used in the studies appear to have a significant impact on performance. For example, larger or more diverse data sets generally provide better generalization ability. This situation has been observed especially in studies using data sets such as FLICKR8K and MSCOCO. More diverse data sets allowed the model to better adapt to different scenarios and produce more universal results. It is important to note that although Transformer-based models are generally more successful, this is not the case for every dataset. For example, studies with high BLEU scores obtained with the VGG16+LSTM model [24] show that different models can be effective in certain situations. As a result, the interaction between the methods and data sets used in studies of producing text from images has a complex structure. Current studies show that different approaches are needed to improve the performance of the system. For this reason, this article shows how the use of the attention mechanism, together with deep learning-based models, increases the performance in text generation.

3. Material and Methods

In this study, image captions were created using the Sequence to Sequence (Seq2seq) model, which is based on the encoder and decoder architecture and incorporates the attention mechanism. The experiment is performed on the Flickr8k dataset, and the file containing the images and their corresponding captions goes through the necessary preprocessing steps. It is desired to measure how much the use of the attention mechanism with a neural network-based model will affect performance. The proposed system with the Attention mechanism was also compared with the VGG19 model. The Seq2seq model architecture comprises an encoder and a decoder. The encoder processes the image features and encodes them into a fixed-size representation. On the other hand, the decoder takes this representation as input and generates captions. The encoder is responsible for extracting image features, and the decoder utilizes these extracted features to create image captions. The inclusion of the Attention Mechanism enhances the decoder's capability to access all hidden states of the encoder, thereby facilitating the generation of more meaningful and comprehensive sentences. The VGG19 model, a remarkable Convolutional Neural Network, was used in this study together with the attention mechanism to increase the overall performance of the text generation process from the image.

3.1 Dataset

Various datasets are utilized in image captioning research. In the literature, datasets such as MS COCO (Microsoft Common Objects in Context), Flickr30k, and Flickr8k are commonly employed for generating English captions. The MS COCO dataset contains high-quality images, consisting of 118,287 images for training and 5,000 images for testing. In total, there are 123,287 images in this dataset, each accompanied by five captions. The Flickr8k dataset consists of 8,091 images typically depicting humans and animals, with corresponding captions for each image. The Flickr30k dataset focuses on images primarily showing people in various events and is an extended version of the Flickr8k dataset. It contains 31,783 images. In this study, the Flickr8k dataset is used. This dataset comprises a total of 8,091 images, with 6,000 for training, 1,000 for testing, and 1,000 for validation. Multiple captions are available for each image in the Flickr8k dataset. Figure 1 shows some tagged images from the Flickr8k dataset.

3.2 Convolutional Neural Network

Convolutional Neural Network (CNN) is a type of neural network that is effective in computer vision tasks, particularly image processing. It aims to extract high-level features using a structure based on n-grams or words, typically used for feature extraction. It finds applications in various domains like image recognition, classification, and text generation. CNN primarily consists of specially designed convolutional layers to process image data [29]. These layers emphasize specific features in the image, aiming to obtain more meaningful and lower-dimensional representations of these features. Pooling layers reduce the dimension of these feature maps while minimizing significant information loss. Activation layers nonlinearly adjust the computed features. The structure of CNN is particularly efficient in extracting and representing features from images (Figure

2). These features can then be used by a decoder for various tasks such as caption generation, classification, or detection. The different layers of convolutional neural networks represent different stages of image processing, while fully connected layers transform these features into actionable results [30].



Figure 1 The used Flickr8k dataset images

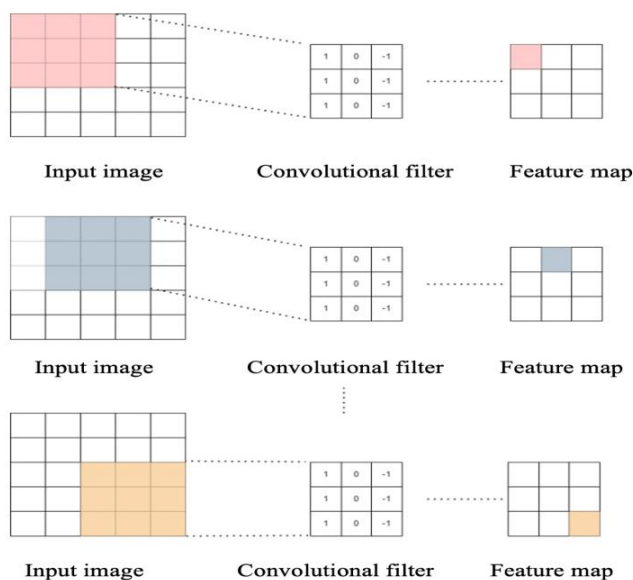


Figure 2 The CNN Architecture Internal Structure

3.3 Sequence-to-Sequence (Seq2Seq) Model

The Sequence-to-Sequence (Seq2seq) model is a deep learning architecture used for tasks such as translating, transforming, or generating sequences of varying lengths between input and output. It consists of two fundamental components: an encoder and a decoder. The encoder converts the input sequence into a fixed-size vector representation [31]. The decoder takes this vector and generates the output sequence. The Seq2seq model has achieved notable success in tasks like language translation, text generation, and speech recognition. Advanced versions incorporating techniques like attention mechanisms have also been developed. In the Seq2seq model, the encoder reads the sequence of input data and passes it to a sequence, which is then fed as input to the decoder to generate the output sequence. As Seq2seq deals with sequential data, both the encoder and decoder typically involve a recurrent structure (RNN, LSTM, GRU). An RNN, for instance, takes into account both the current time step's input and the input from the previous time step. The output at time step "t" is generated based on both the input at time step "t" and the input at time step "t-1". The hidden state in the model retains sequential information and is used in the next step of the process. The encoder takes a sequence as input data. The entire input data is compressed by the encoder into a fixed-size vector, which is then passed to the decoder. The decoder takes this sequence as input. To predict the output data, the decoder progresses from the previous time step's (t-1) hidden state, using the information present until the process is completed. This way, the encoder's hidden state, composed only of the final outputs, struggles less in fully forming sentences [32]. Figure 3 shows the used Seq2seq architecture.

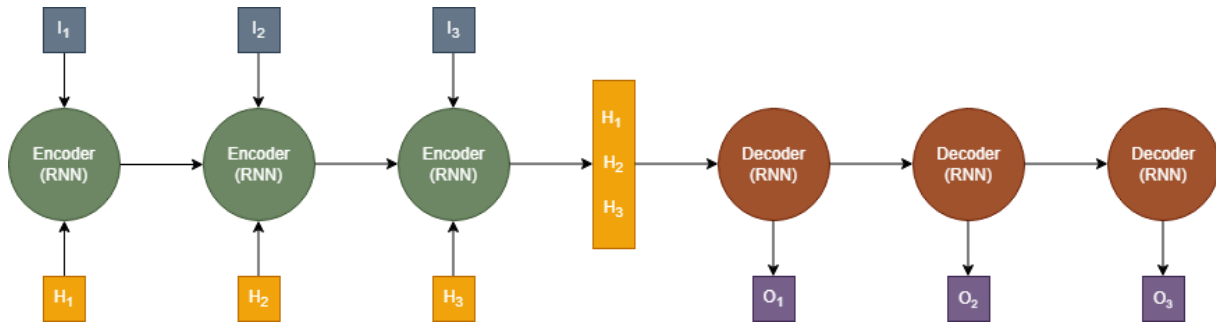


Figure 3 The Used Seq2seq Deep Learning Model Architecture

3.4 Encoder-Decoder

The captions for the input images in the Flickr8k dataset are generated using a Seq2seq model, which is built upon the encoder-decoder architecture. The image is first fed into the encoder. The images are resized to 224x224xRGB dimensions. The encoder extracts features from the image and creates a feature vector. This generated vector is then provided as input to the decoder. As a preprocessing step for the captions, punctuation marks between words are removed, and <s> and <e> tags are added at the beginning and end of each sentence. Tokenization processes are then applied. Using the image feature vector received from the encoder, the decoder constructs the captions.

3.5 Attention Mechanism

An attention mechanism is employed to generate comprehensive captions that encompass all details of the image. The attention mechanism emphasizes relevant information present in the image. This mechanism is based on two fundamental aspects [33]. Firstly, it determines which parts of the input need to be focused on. Secondly, it extracts the features of significant regions. The attention mechanism modifies the context vector at each time step in the decoder based on the similarity of the decoder hidden states. This mechanism is integrated into the encoder-decoder architecture to enhance the generation of image captions with rich details [34].

3.6 VGG19 model

VGG models are CNN models used for image classification. The VGG19 model is named after the Visual Geometry Group and is a model trained using more than 1 million images in the ImageNet database. It is a 19-layer model. The VGG19 model is the VGGNet model with 19 weight layers [35]. It consists of 16 convolutional layers, 3 fully connected layers, 5 max-pool, and 1 softmax layer. The model takes images as input. Once the image is ready for processing, the final image features are extracted.

4. Experimental Results








In text processing studies, different evaluation metrics are employed. BLEU, METEOR, and ROUGE are the most commonly used metrics. These metrics take values between 0 and 100, where a value closer to 100 indicates that the machine translation is similar to the reference translation, while a value closer to 0 suggests that the generated machine translation diverges from the reference translation. The performance evaluation results for Seq2seq and VGG19 models with attention mechanism obtained for text generation on the Flickr8k dataset using the proposed approach are shown in Table 2. Also, Text in English created using images from the Flickr8k dataset is shown in Table 3. Additionally, Table 4 shows a comparison of the proposed approach with existing studies in the literature.

Table 2 Experimental Results for Seq2seq and VGG19 Models

| Evaluation Metrics | Performance of the Seq2seq model | Performance of the VGG19 model |
|--------------------|----------------------------------|--------------------------------|
|--------------------|----------------------------------|--------------------------------|

| | | |
|---------|-------------|--------------|
| BLEU-1 | 71,42 | 66,66 |
| BLEU-2 | 59,76 | 51,63 |
| BLEU-3 | 41,85 | 1,92...e-100 |
| BLEU-4 | 6,31...e-76 | 1,07...e-152 |
| METEOR | 98.13 | 62,5 |
| ROUGE_L | 83,33 | 91,42 |

Table 3. The real captions and prediction captions results

| Image | Captions |
|---|--|
|  | <p>Real Caption: black dog playing with green toy Seq2seq Prediction Caption: black dog plays with a green toy VGG19 Prediction Caption: black dog with a green object</p> |
|  | <p>Real Caption: girl with long hair flying in the breeze while she swings Seq2seq Prediction Caption: The girl in the pink top is swinging with her hair flying everywhere VGG19 Prediction Caption: The girl is swinging with her hair flying everywhere</p> |
|  | <p>Real Caption: A man in a red shirt sits on his dirt bike and points at the camera Seq2seq Prediction Caption: The man with the bike is wearing a helmet is on the bike and pointing at the camera VGG19 Prediction Caption: male bikes through the middle of the mountain</p> |
|  | <p>Real Caption: Real Caption: dirt bike rider jumping down the hill Seq2seq Prediction Caption: A person on a BMX bike is riding on an outdoor course VGG19 Prediction Caption: person rides a vehicle</p> |
|  | <p>Real Caption: The hiker is shadowed by the time of day near an open body of water Seq2seq Prediction Caption: backpacker looks at the ocean sky above the ocean VGG19 Prediction Caption: hiker standing on the shore of the lake</p> |
|  | <p>Real Caption: Young men playing basketball in a competition Seq2seq Prediction Caption: four men playing basketball with the team are in action VGG19 Prediction Caption: A basketball player in white is running with behind him</p> |
|  | <p>Real Caption: an adult and child on bleachers near the water Seq2seq Prediction Caption: A man in a cowboy hat sits on bleachers in the park VGG19 Prediction Caption: Man is sitting on bleachers in front of the lake</p> |

When the performance of the Seq2seq and VGG19 models is evaluated, we observe that the Seq2seq model achieves more impressive results with the effective impression method. While the Seq2seq model exhibited a high similarity with the BLEU-1 score (71.42), it also achieved remarkable success with the METEOR score (98.13) and ROUGE_L score (83.33). On the other hand, the VGG19 model performs slightly lower on similar metrics. Especially in the second image, we see

that the Seq2seq model successfully captures details such as the "pink top." However, he omitted the "red shirt" detail in the third image. In general, the Seq2seq model is better at preserving visual details. The VGG19 model has a more general and abstract perspective. This shows how the attention mechanism, in combination with the Seq2seq model, is more effective, especially in preserving visual details. While the Seq2seq model attracts attention with its ability to integrate visual and textual information more effectively, the VGG19 model focuses on a more general perspective. This study shows that the combination of the Seq2seq model and the attention mechanism can combine visual and textual information in a more meaningful and comprehensive way. More specifically, the attention mechanism of the Seq2seq model tends to preserve visual details better, making it preferable to the VGG19 model.

Table 4 Comparison of Proposed Approach with Other Studies

| Author | Dataset | Method | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE_L |
|--------------------------------|----------|------------------|--------|--------|--------------|--------------|--------|---------|
| Singh et al. [13] | FLICKR8K | CNN+LSTM | 0.52 | -- | -- | -- | -- | -- |
| Rafi et al. [17] | FLICKR8K | InceptionV3+LSTM | 81.1 | 38.3 | -- | -- | -- | -- |
| Kushwaha et al. [24] | FLICKR8K | VGG16+LSTM | 55.9 | -- | -- | -- | -- | -- |
| Proposed approach with Seq2seq | FLICKR8K | Seq2seq | 71,42 | 59,76 | 41,85 | 6,31...e-76 | 98.13 | 83,33 |
| Proposed approach with VGG19 | FLICKR8K | VGG19 | 66,66 | 51,63 | 1,92...e-100 | 1,07...e-152 | 62,5 | 91,42 |

In the study, Seq2seq and VGG19 models, evaluated with common metrics such as BLEU, METEOR, and ROUGE, were used to measure the performance of the proposed approach. Looking at Table 4, the Seq2seq model has higher values in BLEU-1 and BLEU-2 metrics compared to other studies. Especially in BLEU-1, it achieved a success of 71.42%. This indicates that the Seq2seq model contributes to a better matching of words under the proposed approach than similar studies in previous literature. In the ROUGE_L metric, the VGG19 model showed a higher performance compared to other studies (91.42%). This may indicate that VGG19, enhanced by the attention mechanism, has better similarity with reference texts. When we look at the comparisons in Table 4, it can be seen that the proposed approach, Seq2seq+Attention mechanism, has a competitive advantage compared to previous studies. However, the VGG19 model performed poorly on certain metrics, especially when compared to other models. The use of the attention mechanism showed a particularly pronounced effect on the Seq2seq model. The Seq2seq model achieved better results when supported by the attention mechanism and stood out, especially in BLEU-1 and BLEU-2 metrics. This indicates that the attention mechanism helps in making better word alignments under the proposed approach in the text generation task.

5. Conclusion

This study aims to make a substantial contribution to the existing literature by thoroughly investigating the integration of deep neural network models and the utilization of the attention mechanism in the realm of image-to-text conversion. The proposed approach, implemented in conjunction with the Seq2seq model, renowned for its proficiency in sequential data transformation, and VGG19, a widely adopted convolutional neural network model, was subjected to rigorous comparisons with other studies in the literature. Our experiments, conducted on the Flickr8k dataset, underscore the exceptional capability of our proposed approach in generating captions that closely align with reference sentences. The dynamic focusing facilitated by the attention mechanism on various parts of the images enhances the captions by capturing intricate details. Future endeavors will leverage larger and more diverse datasets, delve into advanced attention mechanisms, and explore transfer learning and fine-tuning techniques to enhance adaptability across different domains. Our findings emphasize the potential of our approach in visual understanding, content creation, and human-computer interaction. The proficient integration of visual and textual information positions our model as a valuable asset in the ever-evolving landscape of multimodal data processing. With the continuous advancements in deep learning, our proposed approach holds the promise of pushing the boundaries of efficient multimodal data representation and processing, underlining the superiority of Seq2seq coupled with the attention mechanism in achieving compelling results.

References

- [1] T. Alqahtani et al., "The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research," *Research in Social and Administrative Pharmacy*, vol. 19, no. 8, pp. 1236–1242, Aug. 2023, doi: 10.1016/j.sapharm.2023.05.016.
- [2] J. J. Cavallo, I. de Oliveira Santo, J. L. Mezrich, and H. P. Forman, "Clinical Implementation of a Combined Artificial Intelligence and Natural Language Processing Quality Assurance Program for Pulmonary Nodule Detection in the Emergency Department Setting", *Journal of the American College of Radiology*, vol. 20, no. 4, pp. 438–445, Apr. 2023, doi: 10.1016/j.jacr.2022.12.016.
- [3] J. Doe and A. Smith, "Recent advances in image captioning: A comprehensive survey," *IEEE Transactions on Artificial Intelligence*, vol. 7, no. 3, pp. 210-225, 2022.
- [4] M. Johnson, B. Brown, and C. Wilson, "Innovative Approaches for image caption generation using attention mechanisms," *Proceedings of the 35th Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, 2021, pp. 750-760.
- [5] S. Kim and E. Lee, "Enhancing image captioning performance through multimodal fusion techniques," *IEEE Transactions on Multimedia*, vol. 25, no. 6, pp. 1350-1365, 2020.
- [6] L. Wang, H. Chen, and X. Zhang, "Leveraging transformers for improved image captioning," *Proceedings of the European Conference on Computer Vision (ECCV)*, Munich, Germany, 2018, pp. 240-255.
- [7] R. Patel and S. Gupta, "Attention is all you need: Exploring self-attention mechanisms in image captioning," *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, HI, USA, 2019, pp. 1800-1810.
- [8] C. Yue, W. Hu, H. Song, and W. Kang, "Thangka image caption method based on attention mechanism and encoder-decoder architecture" In *2022 7th International Conference on Intelligent Computing and Signal Processing (ICSP)*, Nis. 2022, pp. 1752-1756. doi: 10.1109/ICSP54964.2022.9778737.
- [9] S. R. Chandaran, S. Natesan, G. Muthusamy, P. K. Sivakumar, P. Mohanraj, and R. J. Gnanaprakasam, "Image captioning using deep learning techniques for partially impaired people" In *2023 International Conference on Computer Communication and Informatics (ICCCI)*, Oca. 2023, pp. 1-6. doi: 10.1109/ICCCI56745.2023.10128287.
- [10] S. S. Bhadauria, D. Bisht, T. Poongodi, and S. A. Yadav, "Assertive vision using deep learning and LSTM" In *2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM)*, Şub. 2022, pp. 761-764. doi: 10.1109/ICIPTM54933.2022.9754057.
- [11] M. K. Shaikh and M. V. Joshi, "Recursive network with explicit neighbor connection for image captioning" In *2018 International Conference on Signal Processing and Communications (SPCOM)*, Tem. 2018, pp. 392-396. doi: 10.1109/SPCOM.2018.8724400.
- [12] Z. Xue, L. Wang, and P. Guo, "Slot based image captioning with WGAN" In *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*, Haz. 2019, pp. 241-246. doi: 10.1109/ICIS46139.2019.8940218.
- [13] A. Singh et al., "Image captioning using python" In *2023 International Conference on Power, Instrumentation, Energy and Control (PIECON)*, Şub. 2023, pp. 1-5. doi: 10.1109/PIECON56912.2023.10085724.
- [14] S.-H. Han and H.-J. Choi, "Explainable image caption generator using attention and Bayesian inference" In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, Ara. 2018, pp. 478-481. doi: 10.1109/CSCI46756.2018.00098.
- [15] B. Wang et al., "Cross-lingual image caption generation based on visual attention model" *IEEE Access*, vol. 8, pp. 104543-104554, 2020, doi: 10.1109/ACCESS.2020.2999568.
- [16] H. Chen et al., "Captioning transformer with scene graph guiding" In *2021 IEEE International Conference on Image Processing (ICIP)*, Eyl. 2021, pp. 2538-2542. doi: 10.1109/ICIP42928.2021.9506193.
- [17] S. Rafi and R. Das, "A linear sub-structure with co-variance shift for image captioning" In *2021 8th International Conference on Soft Computing & Machine Intelligence (ISCFMI)*, Kas. 2021, pp. 242-246. doi: 10.1109/ISCFMI53840.2021.9654828.
- [18] D. Sharma et al., "gntweight transformer with GRU integrated decoder for image captioning" In *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Eki. 2022, pp. 434-438. doi: 10.1109/SITIS57111.2022.00072.
- [19] L. Panigrahi et al., "Hybrid image captioning model" In *2022 OPJU International Technology Conference on Emerging*

- Technologies for Sustainable Development (OTCON)*, Şub. 2023, pp. 1-6. doi: 10.1109/OTCON56053.2023.10113957.
- [20] Q. Sun et al., “Bidirectional Beam Search: Forward-Backward inference in neural sequence models for fill-in-the-blank image captioning” In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Tem. 2017, pp. 7215-7223. doi: 10.1109/CVPR.2017.763.
- [21] Y. Keneshloo et al., “Deep reinforcement learning for sequence-to-sequence models”, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 7, pp. 2469-2489, Tem. 2020, doi: 10.1109/TNNLS.2019.2929141.
- [22] M. Sahrial Alam et al., “Arison of different CNN model used as encoders for image captioning” In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)*, Eki. 2021, pp. 523-526. doi: 10.1109/ICDABI53623.2021.9655846.
- [23] Y. Zhou et al., “Triple sequence generative adversarial nets for unsupervised image captioning” In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Haz. 2021, pp. 7598-7602. doi: 10.1109/ICASSP39728.2021.9414335.
- [24] R. Kushwaha and A. Biswas, “Hybrid feature and sequence extractor based deep learning model for image caption generation” In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Tem. 2021, pp. 1-6. doi: 10.1109/ICCCNT51525.2021.9579897.
- [25] H. Liu et al., “Vocabulary-wide credit assignment for training image captioning models” *IEEE Transactions on Image Processing*, vol. 30, pp. 2450-2460, 2021, doi: 10.1109/TIP.2021.3051476.
- [26] Y. Zheng et al., “Divcon: Learning concept sequences for semantically diverse image captioning” In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Haz. 2023, pp. 1-5. doi: 10.1109/ICASSP49357.2023.10094565.
- [27] B. Birmingham and A. Muscat, “KENGIC: Keyword-driven and N-Gram graph based image captioning” In *2022 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Nov. 2022, pp. 1-8. doi: 10.1109/DICTA56598.2022.10034584.
- [28] P. Dwivedi and A. Upadhyaya, “A Novel deep learning model for accurate prediction of image captions in fashion industry” In *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Oca. 2022, pp. 207-212. doi: 10.1109/Confluence52989.2022.9734171.
- [29] F. Akalin and N. Yumusak, "Detection and classification of white blood cells with an improved deep learning-based approach," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 30, no. 7, article 16. <https://doi.org/10.55730/1300-0632.3965>
- [30] F. Akalin, and N. Yumusak. "Classification of ALL, AML and MLL leukaemia types on microarray dataset using LSTM neural network Approach" , *Journal of Faculty of Engineering and Architecture of Gazi University*, vol. 38, no. 3, 2023, pp. 1299-1306.
- [31] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks” in *Advances in Neural Information Processing Systems*, (pp. 3104-3112), 2014.
- [32] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 6, pp. 1416-1432, 2018.
- [33] J. Xie et al., “A multimodal fusion emotion recognition method based on multitask learning and attention mechanism”, *Neurocomputing*, p. 126649, Aug. 2023, doi: 10.1016/j.neucom.2023.126649.
- [34] K. Yang et al., “A multi-sensor mapping Bi-LSTM model of bridge monitoring data based on spatial-temporal attention mechanism”, *Measurement*, vol. 217, p. 113053, Aug. 2023, doi: 10.1016/j.measurement.2023.113053.
- [35] H. Won, B. Kim, I.-Y. Kwak, ve C. Lim, “Using various pre-trained models for audio feature extraction in automated audio captioning,” *Expert Systems with Applications*, c. 231, s. 120664, Kas. 2023, doi: 10.1016/j.eswa.2023.120664.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Availability of Data and Material

Not applicable.

Ethical Approval and Informed Consent

It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Plagiarism Statement

This article has been scanned by iThenticate TM.