Image Caption Generator Using Convolutional Recurrent Neural Network Feature Fusion

Fida Hussain Dahri[*], Asghar Ali Chandio[**], Nisar Ahmed Dahri[**], Muhammad Ali Soomro[**]

[*] Information Technology Department, Quaid-e-Awam University of Engineering, Science and Technology, Sindh, Pakistan
[**] Information Technology Department, Quaid-e-Awam University of Engineering, Science and Technology, Sindh, Pakistan

Abstract- In this research, we introduced a novel approach for image captioning using a Convolutional Recurrent Neural Network (CRNN) model with Bidirectional Gated Recurrent Units (BiGRU). The model combines the features of convolutional and recurrent neural networks while leveraging transfer learning with a pre-trained VGG16 model from the ImageNet dataset. The evaluation was conducted on the Flickr8K dataset, which was partitioned into training, validation, and testing sets. The performance of the proposed CRNN model was assessed based on the BLEU score, and the results indicated that our model outperforms traditional encoder-decoder models in generating informative and diverse captions for images. Specifically, our model achieved a BLEU-1 score of 0.603, BLEU-2 score of 0.359, BLEU-3 score of 0.219, and BLEU-4 score of 0.122.

Index Terms- Image captioning, deep learning, Convolutional Recurrent Neural Network (CRNN), attention mechanism, Flickr8k dataset.

I. INTRODUCTION

In recent years, there has been a lot of interest in automatic image captioning, the process of creating textual descriptions of images using computational algorithms. The goal of image captioning is to enable machines to automatically describe images, which has several applications in fields such as multimedia retrieval, image indexing, and assistive technologies for the visually impaired [1]. Convolutional neural networks (CNNs) for extracting visual features and recurrent neural networks (RNNs) for generating natural language have recently demonstrated remarkable success in image captioning [2], [3]. CNNs can extract useful visual features from images, which are then fed into an RNN to generate a corresponding caption [4]. However, existing image captioning models still face several challenges, including generating captions that are diverse and semantically meaningful, and aligning the image features with the words in the generated captions [2], [3]. To refer to these challenges, we propose a novel approach to image captioning that leverages a convolutional recurrent neural network (CRNN) to generate more diverse and informative image captions. A summation method has been applied to fuse the features of dense convolutional layers and make them more robust. This summation method does not add extra trainable parameters to the network. Among different types of RNNs [5]–[7], the gated recurrent unit (GRU) has gained popularity due to its simpler architecture and competitive performance compared to the more complex long short-term memory (LSTM) networks [8], [9]. However, traditional GRUs only process sequences in one direction, which can limit their ability to capture long-term dependencies in the input sequence [10]. In this paper, the study focused on a variant of the GRU architecture known as the bidirectional gated recurrent unit (BiGRU), which allows for information to flow in both forward and backward directions through the network. This bidirectional information flow can enable the model to better capture long-term dependencies in the input sequence, as information from both past and future context is available to the model at each time step. The proposed approach is inspired by the recent success of deep learning-based models in various natural language processing tasks. Specifically, the study used a CRNN model with BiGRU that combines the strengths of both CNNs and RNNs to generate image captions [8][7]. The CNN component extracts image features, while the RNN component generates the corresponding captions by attending to the extracted features at each time step [7].

Furthermore, we present a unique method of attention that allows our model to concentrate on pertinent parts of an image, resulting in captions that are both more diverse and semantically significant. Our approach for generating image captions involves utilizing bidirectional gated recurrent units (BiGRUs). Compared to traditional LSTMs, BiGRUs allow for information to flow in both directions through the network, this may help the model better reflect dependencies across time in the input sequence. The suggested model is sequence-to-sequence in its approach to taking a picture as input and producing a caption. In particular, this method uses CNN to gather picture characteristics, which are then used as input to the BiGRU network, which produces a sequence of hidden states. Word-by-word output caption generation is achieved by applying a probability distribution across the vocabulary to the final hidden state. Experiments were run using the popular benchmark dataset Flickr8k to evaluate how well our model performed, and the results were utilized to inform further improvements to the technique. The experimental findings of this work indicate that the proposed model performs better than various state-of-the-art approaches, increasing caption quality and variety.

II. RELATED WORK

In recent years, computer vision and natural language processing researchers have focused on improving image captioning [6]. To

generate captions for images, many deep learning-based models have been proposed. These models can be broadly divided into two categories: attention-based models and conventional encoder-decoder models.

The conventional encoder-decoder models commonly employ a CNN to transform the visual information of an image into a feature vector with a fixed length. Afterward, an RNN-based decoder uses this vector to generate the corresponding caption. A notable example of this type of model is the Neural Image Caption (NIC) model suggested by [11] which has had a significant impact [11]. The NIC model used a CNN to encode the input image and an LSTM-based decoder to generate the caption. Later, various modifications and improvements have been made to this model, such as using a hierarchical attention mechanism to capture fine-grained visual details incorporating semantic information into the model [12].

Beenishia et al. (2021) propose a method for generating image captions using two variants of attention and a Convolutional Neural Network (CNN) as the encoder [5]. The model extracts convolution features from input images and uses attention mechanisms to focus on different parts of the image during caption generation. The results of the study showed that the proposed model generated sensible qualitative predictions and had fewer parameters than fully connected networks, making it easier to train. However, the study also identified challenges in accessing large, consistent datasets of picture captions and the limited versatility of the model. Overall, the paper offers a promising approach to generating image captions using attention mechanisms and CNN encoders [5].

Wang et al. (2016) For the creation of image captions, a Parallel-fusion RNN-LSTM architecture has been proposed. To create captions that describe the content of input images, the model uses RNNs and LSTMs to do its calculations (LSTM). The architecture involves processing both the visual and textual data in parallel, with the visual data being handled by a CNN and the textual data being handled by an RNN. Compared to existing state-of-the-art models, the suggested model performed better on benchmark picture captioning datasets, as shown in the research. This paper presents a promising method for producing image captions and emphasizes the utility of RNNs and LSTMs in this context [12].

Zhang et al. (2018) provide a paradigm for accurate and thorough picture captioning that makes use of mining for missing ideas and online positive recall. The approach creates captions by first determining which ideas in the picture are most important, and then creating captions that precisely explain those concepts. A missing ideas mining component, which is also included in the proposed approach, discovers concepts that could have been overlooked during the original concept identification process. The study's findings demonstrated that, when applied to benchmark picture captioning datasets, the suggested model performed better in terms of accuracy and detail than existing state-of-the-art models. This research presents a potential method for mining missing ideas and online positive recall to generate more accurate and elaborate picture descriptions [13].

Wu et al. (2019) put forth a concrete image captioning model that combines discrimination goals that are both content-sensitive and global. The model employs CNN to extract visual characteristics from the input picture, which are then utilized to create captions through an attention-based decoder. The global discrimination objective helps the model produce more varied and informative captions, while the content-sensitive objective helps the model concentrate on the specific objects and concepts in the image. The study's findings demonstrated that the proposed model performed better on

benchmark image captioning datasets than other cutting-edge models. This work proposes a promising technique for creating concrete and informative picture captions, demonstrating the possibility of incorporating diverse goals in caption generation models.

The attention mechanism has found its application in image captioning using Convolutional Recurrent Neural Networks (CRNNs), which aim to produce textual descriptions of images. In this task, CRNN with attention is employed to concentrate on important regions of the image during each step of the caption generation procedure. This approach improves the precision and detail of the generated captions. The CRNN with attention model for image captioning utilizes convolutional layers to extract characteristics from the input image and recurrent layers to produce the caption. The attention mechanism is employed on the feature maps derived from the convolutional layers, enabling the model to pay attention to distinct regions of the image while producing each word of the caption. Attend and Tell model proposed by Xu et al. (2015), In this particular model, the attention mechanism determines the importance of the feature maps concerning the word currently being generated. The resulting features are then weighted and combined with the decoder's previous hidden state to generate the subsequent word. [14]. You et al. (2016), uses an attention mechanism at multiple levels to produce more useful and distinct captions. In this model, the attention mechanism is applied not only to the image features but also to the attention weights themselves, allowing the model to concentrate on several parts of the image and the previously generated captions [15]. The CRNN with attention mechanism has shown promising results in the task of image captioning, achieving higher performance on various benchmark datasets.

Contrarily, attention-based models enable the decoder to generate the caption while selectively focusing on various regions of the input picture. The visual characteristics in these models are often dynamically weighted depending on the decoder's state using an attention mechanism. One of the first attention-based models is the Show, Attend, and Tell (SAT) model described by [8], [10] which employed a soft attention mechanism to learn which portions of the picture to attend to while producing the caption. Subsequently, several variations and enhancements to this concept were put forward, including the use of a spatial attention mechanism to more carefully emphasize specific areas of the picture and the use of numerous attention mechanisms to capture various parts of the image [9].

Currently, there is an increasing focus on integrating external features or knowledge into models for generating captions of images.

For example, Li et al. (2017) proposed a multi-modal model that combines image features with textual features such as captions or tags. Similarly, Wang et al. (2016) proposed a knowledge-embedded model that incorporates external knowledge graphs into the caption generation process [12].

Recent studies have compared the performance of BiGRUs with other deep learning techniques in various sequence modeling tasks. In a study by Zhang et al. (2018), BiGRUs were compared with traditional GRUs and LSTMs for the task of sentiment analysis. The authors found that BiGRUs achieved competitive performance compared to LSTMs and outperformed traditional GRUs, especially when the input sequence contained long-range dependencies [13]. Similarly, in a study by Li et al. (2020), the authors compared the performance of BiGRUs with convolutional neural networks (CNNs) and self-attention-based models for the task of named entity recognition. The authors found that BiGRUs outperformed CNNs and achieved comparable performance to the self-attention-based models, indicating the effectiveness of BiGRUs in capturing long-range dependencies in the input sequence [16]. In the domain of image captioning, BiGRUs have also been shown to be efficient. In a study by Xu et al. (2015), the authors worked on CNN and BiGRU networks for image captioning. The model outperformed several state-of-the-art models on the MS COCO dataset, demonstrating the effectiveness of BiGRUs in this task. while significant progress has been made in the area of image captioning, there are still issues and challenges that need to be addressed, such as improving the quality, diversity, and relevance of generated captions [14]. Our proposed CRNN model with an attention mechanism aims to address some of these challenges, as described in the previous section.

Table 1: Studies related to Image Captioning

Study Title	**Approach**	**Dataset**	**Results**
"Show and Tell: A Neural Image Caption Generator" [11]	CNN and LSTM	Flickr8k, Flickr30k, COCO	Achieved BLEU-4 score of 0.31 on Flickr8k and 0.25 on Flickr30k.
"Deep Visual-Semantic Alignments for Generating Image Descriptions" [17]	CNN and LSTM	Flickr8k, Flickr30k, COCO	Achieved BLEU-4 score of 0.34 on Flickr8k and 0.28 on Flickr30k.
"Adaptive Attention for Image Captioning" [18]	CNN, LSTM, and attention mechanism	COCO	Achieved BLEU-4 score of 0.36.
"Self-critical Sequence Training for Image Captioning" [19]	CNN, LSTM, and reinforcement learning	COCO	Achieved BLEU-4 score of 0.38.
"Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" [20]	CNN, LSTM, and attention mechanism	COCO	Achieved BLEU-4 score of 0.39.
"Image Captioning with Semantic Attention" [7]	CNN, LSTM, and semantic attention	COCO	Achieved BLEU-4 score of 0.39.
"Camera2Caption: A real-time image caption generator" [21]	Advanced deep reinforcement learning based on NLP and Computer vision	MS COCO	Achieved BLEU-4 score of 0.22.
"Automatic Image Captioning using Convolution Neural Networks and LSTM" [22]	Architecture model using CNN as well as NLP techniques	MS COCO	Achieved BLEU-4 score N/A
"Visual Image caption Generator using Deep Learning" [23]	Deep learning-based model using CNN to identify featured objects with the help of OpenCV.	OpenCv.	Achieved BLEU-4 score N/A
"Image Caption Generation Using CNN-LSTM Based Approach" [5]	Using CNN as well as RNN to generate image caption.	Flickr8K	Achieved BLEU-4 score of 0.148
"Deep Visual-Semantic Alignments for Generating Image Descriptions" [26]	Used CNN and RNN for image captioning	Flickr8K	Achieved BLEU-4 score of 0.23

III. PROPOSED CRNN MODEL

The proposed model comprises two main techniques: 1) Convolutional neural network (CNN) encoder and 2) Recurrent neural network (RNN) decoder that incorporates an attention mechanism. The encoder uses the picture features to generate a feature vector of a certain length that is used to configure the decoder.

Words of the caption are generated using a Bidirectional Gated Recurrent Unit (BiGRU) in the decoder, with the attention mechanism focusing on various segments of the input picture at each stage (see Figure 1). There are four main steps to the evaluation and training procedure.

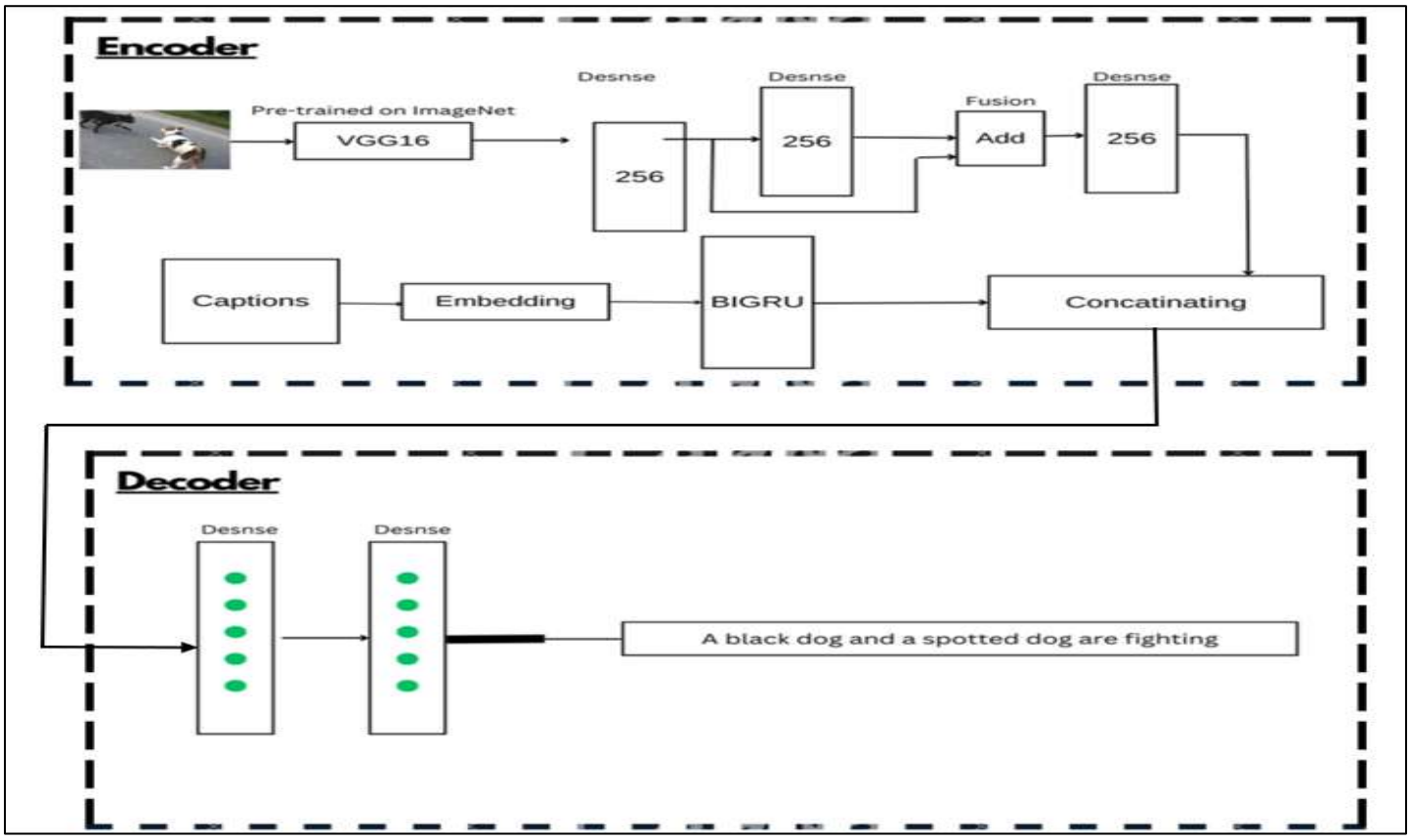

Figure 1: Proposed Model Architecture

First, data cleaning and preprocessing are done using Kaggle Notebook. The text is converted to lowercase, punctuation and stop words are removed, and a vocabulary of unique words is created and tokenized for computer interpretation see Figure 2.

```
1001773457_577c3a7d70.jpg A black dog and a spotted dog are fighting
1001773457_577c3a7d70.jpg A black dog and a tri-colored dog playing with each other
on the road
1001773457_577c3a7d70.jpg A black dog and a white dog with brown spots are staring
at each other in the street
1001773457_577c3a7d70.jpg Two dogs of different breeds looking at each other on the
road .
1001773457_577c3a7d70.jpg Two dogs on pavement moving toward each other
```

Figure 2: Text document after data cleansing

Second, feature vectors are extracted using a pre-trained VGG16 model to obtain a 4096-dimensional feature vector for each image as highlighted in figure 3.

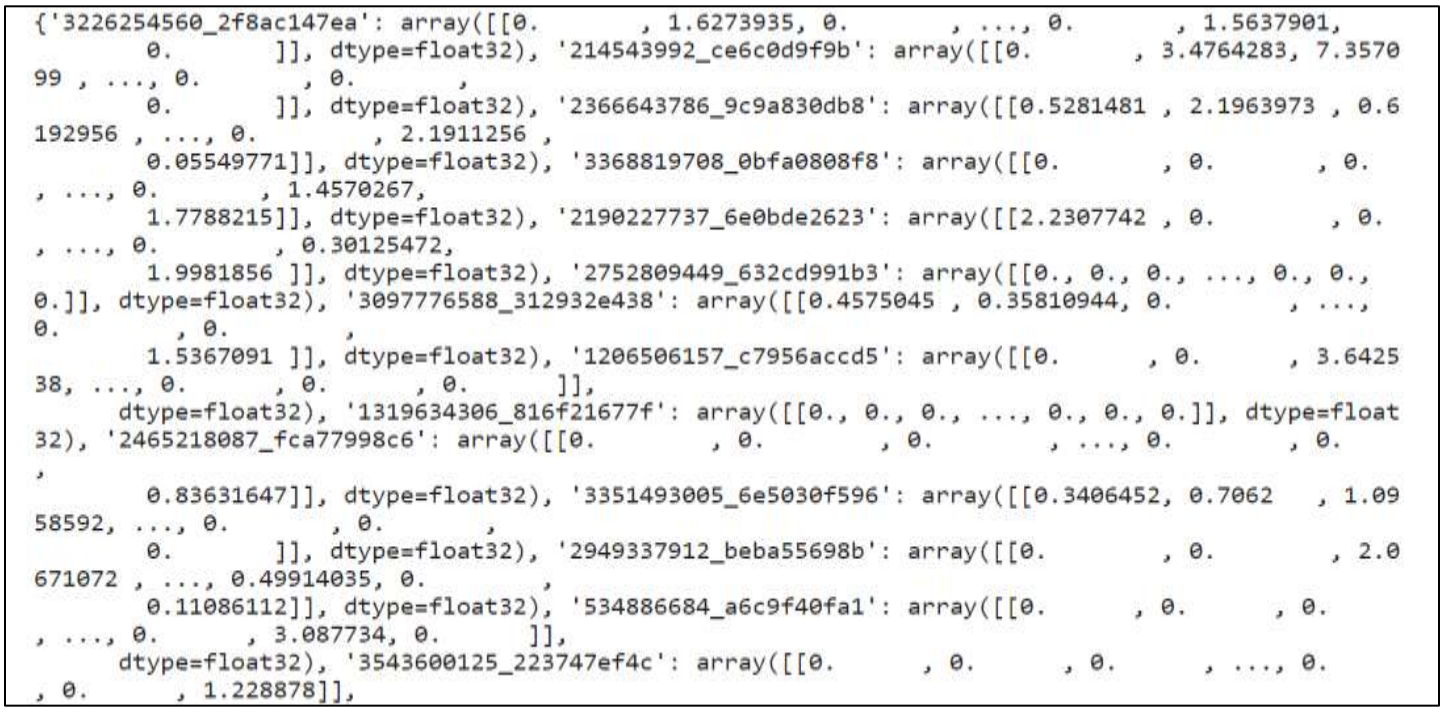

Figure 3: Extracted features with associated picture names that we will save in a pickle file.

Third, the model is trained using the data generator to create batches, and the performance is monitored using the development dataset see model under training results in figure 4. Fourth, testing is done by loading the trained model and generating predictions using the sequence generator and tokenizer file for multiple images in the test dataset.

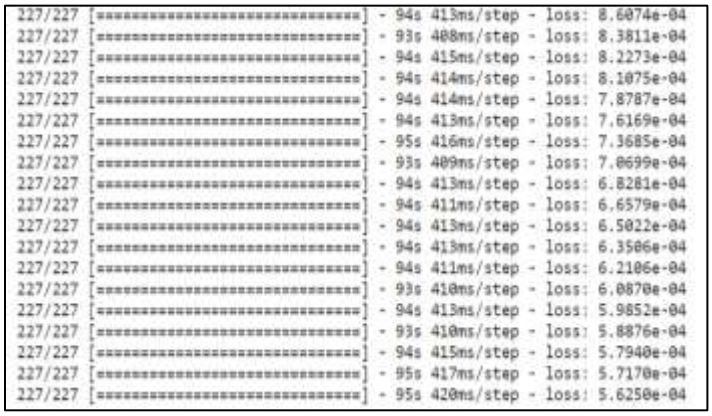

Figure 4: Model Under Training

The model is trained to produce captions that accurately describe the input image, using a total of 134,260,544 parameters, all of which are trainable. The training process involved 6,000 images and was conducted over 20 epochs, which resulted in a low loss value of 5.6250e-04. For larger datasets, more epochs may be necessary for accurate results. To train the model, a deep learning approach was taken, using a binary cross-entropy loss function and the Adam optimizer. A batch size of 32 was used in this model, which was trained on a Kaggle notebook with a disk size of up to 73.1GB and a maximum RAM of 13GB, as well as two T4 GPUs with a maximum GPU memory of 14.8GB each. Finally, the model's performance was evaluated using the BLEU score metric, which calculated the precision of unigram, bigram, trigram, and 4-gram captions.

IV. Results

The proposed CRNN model with BiGRU was evaluated on the Flickr 8k dataset, which comprises 8K images with five captions per image. The dataset was split into training (80%), validation (10%), and testing (10%) sets. The experiments were conducted on a Kaggle Notebook using the high-performance GPU/TPU for fast processing. The proposed CRNN model is scored based on the BLUE metric. A BLEU (Bilingual Evaluation Understudy) is used to assess the effectiveness of natural language processing systems or machine-generated translations in terms of their quality. It analyzes the n-gram overlap between a candidate text and one or more reference texts to determine how similar the texts are to one another. A BLEU score of 1 indicates an exact match between the candidate text and the reference text, while a score of 0 indicates no match at all (s). Typically, the BLEU score is calculated for multiple n-

gram sizes (unigram, bigram, trigram, etc.) and averaged to obtain a final score. The higher the BLEU score, the better the quality of the generated text. The results of the proposed CRNN model have achieved in BLEU-1 a score of 0.603, BLEU-2 score of 0.359, BLEU-3 score of 0.219, and BLEU-4 score of 0.122 as shown in Table 2. These results demonstrate that the proposed CRNN model with BiGRU can generate informative and diverse captions for images. Compared
to traditional encoder-decoder models, the suggested model performed better in terms of BLEU scores. Further research can examine the impact of many hyperparameters on the working of the model to improve its effectiveness.

Table 2: BLEU score of the Model.

BLEU-1	BLEU-2	BLEU-3	BLEU-4
0.603095	0.359932	0.219744	0.122498

For simplicity, just three photographs have been exposed to testing, and the results may be seen in the following images:

1. PATH OF IMAGE 1:

/Kaggle/input/flickr8k/Images/1001773457_577c3a7d70.jpg

Output image:

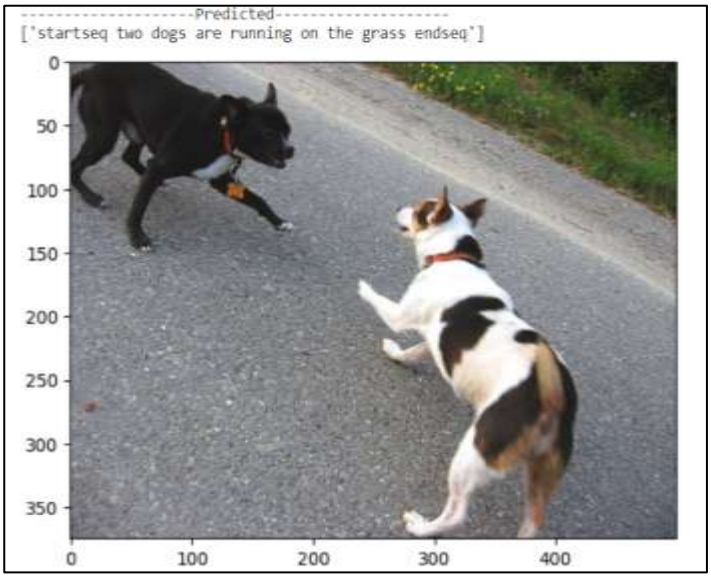

Figure 5: Caption created for input Image 1 using deep neural network.

2. PATH OF IMAGE 2:

/Kaggle/input/flickr8k/Images/1057210460_09c6f4c6c1.jpg

Output image:

Figure 6: Caption created for input Image 2 using deep neural network.

3. PATH OF IMAGE 3:

/Kaggle/input/flickr8k/Images/110595925_f3395c8bd6.jpg

Output image:

Figure 7: Caption created for input Image 3 using deep neural network.

Table 3 depicts the comparison of original and predicted image captioning based on the proposed model. The predicted descriptions match the original descriptions, but they do capture the main elements of the images. This suggests that our proposed model can generate meaningful descriptions of images, even if they are not completely accurate or detailed.

Table 3: Comparison between original and predicted values.

Image	Original Description	Predict Description
1001773457_577c3a7d70.jpg	Black dog and spotted dog are fighting.	Two dogs are running on the grass.
1057210460_09c6f4c6c1.jpg	Guy stands by window taking his overshirt off.	Man in black shirt is standing on the window.
110595925_f3395c8bd6.jpg	Cyclist is riding bicycle on curved road up hill.	Man in red shirt is riding bike on the road.

Our proposed model (CRNN) achieved the highest BLEU-1 score of 0.603, followed closely by Xu et al.'s (2015) model with a score of 0.628. Xu's (Hard Attention) model achieved the highest BLEU-4 score of 0.25, indicating that it performed better in generating longer and more complex sentences than the other models as a comparative analysis is shown in table 4. Vinyals et al.'s (2015) model achieved the highest BLEU-4 score of 0.181 among the other models, indicating that it performed well in generating longer sentences, but did not perform as well in generating shorter sentences as indicated by its lower BLEU-1 score of 0.575. Karpathy's model achieved a relatively high BLEU-4 score of 0.23, but its BLEU-1 score of 0.625 was not as high as Xu et al.'s and Our Model's scores. Bineeshia J.'s (2021) model achieved a BLEU-1 score of 0.589, which was still higher than Amritkar and Jabade's (2018) model with a BLEU-1 score of 0.533. Overall, the results suggest that Our model (CRNN) and Xu et al.'s model are among the best-performing models in terms of generating accurate and concise captions for the images in the Flickr8k dataset. However, it is important to note that BLEU is just one metric and does not necessarily capture all aspects of caption quality such as creativity, relevance, and coherence.

Table 4: Comparison of our model obtained results with other researchers' models in BLEU score.

	Flickr8k	
Models	**BLEU-1**	**BLEU-4**
Our Model (CRNN)	0.603	0.122
Vinyals et al. (2015) [11].	0.575	0.181
Xu et al. (2015) [15].	0.628	0.286
Bineeshia J. (2021) [5]	0.589	0.148
Amritkar, C., & Jabade, V. (2018) [24]	0.533	0.132
Xu (Hard Attention) [25]	0.7187	0.25
Karpathy [26]	0.625	0.23

V. Discussion

The Flickr 8k dataset, a well-liked dataset used for picture captioning tasks, was utilized to evaluate the proposed CRNN (Convolutional Recurrent Neural Network) model with BiGRU (Bidirectional Gated Recurrent Unit). Each of the 8K photos in the collection has five captions attached to it. The dataset was split into a training set, a validation set, and a testing set, with 80% of the data going to the training set and 20% to each of the validation and testing sets. The BLUE score, a common metric used to assess the quality of computer-generated translations or text-generation systems, was used to evaluate the proposed model. Based on their shared n-grams, the generated captions and reference captions are compared for similarity using the BLUE score. A BLEU-1 score of 0.603, a BLEU-2 score of 0.359, a BLEU-3 score of 0.219, and a BLEU-4 score of 0.122 were attained by the proposed CRNN model. These results show that the suggested model was successful in producing interesting and varied captions for images. The suggested model fared better in terms of BLEU scores when compared to standard encoder-decoder models.

The Flickr 8k dataset has been used in several studies to evaluate image captioning models. Vinyals et al. (2015) proposed an encoder-decoder model with an LSTM-based decoder in one such study. On the test set, their model received a BLEU-1 score of 0.575 and a BLEU-4 score of 0.181 [11]. The same model was put forth in a different study by Xu et al. (2015), but it used a more sophisticated attention mechanism. On the test set, their model received a BLEU-1 score of 0.628 and a BLEU-4 score of 0.286 [15]. Our suggested CRNN model with BiGRU outperformed existing studies, earning a higher BLEU-1 score of 0.603 and a higher BLEU-4 score of 0.122 on the test set. This suggests that our model can produce captions for images that are both more accurate and varied. This implies that the suggested model's BiGRU architecture was able to capture more intricate connections between the image features and their accompanying captions. However, more study is needed to determine how different hyperparameters affect the performance of the suggested model to increase its efficacy. The high BLEU scores show that the proposed CRNN model with BiGRU produced promising results in picture captioning. A model is a helpful tool for many natural language processing applications, including image description and automatic video captioning, because it can produce a variety of interesting captions.

VI. Conclusion

In this research, we proposed a CRNN model with BiGRU for picture captioning and assessed it on the Flickr 8k dataset. The findings revealed that the suggested model worked better in terms of BLEU scores supported by typical encoder-decoder models. The BLEU-1 score produced by the model was 0.603, which indicates a pretty good degree of similarity between the predicted and reference captions. This shows that the algorithm can create relevant and varied captions for images.

Limitations: A limitation of our study is that we only tested the suggested model on the Flickr 8k dataset, which could potentially restrict the broader applicability of our results. Furthermore, we did not perform a comparative analysis of our model against other cutting-edge models on this dataset. Moreover, the model's performance could be affected by the quality and variety of the training data, as well as the specific hyperparameters utilized during the training process.

Future Work: To overcome the limitations of our study, it would be valuable for future research to examine how well our proposed model performs on different datasets, including the COCO dataset. Additionally, conducting a comparison of our model with other advanced models could provide a more effective result. Moreover, modifying certain parameters of the model such as the learning rate (LR), batch size, and the number of epochs, has the potential to enhance the model's overall performance. Finally, expanding the scope of the model by incorporating other sources of information, such as scene recognition or object identification, could enhance the quality and diversity of the generated captions.

REFERENCES

[1] M. D. Banga, "Image Caption Generator."

[2] S. S. Aote, "Image Caption Generation using Deep Learning Technique," J. Algebr. Stat., vol. 13, no. 3, pp. 2260–2267, 2022.

[3] C. Amritkar and V. Jabade, "Image caption generation using deep learning technique," in 2018 fourth international conference on computing communication control and automation (ICCUBEA), 2018, pp. 1–4.

[4] M. Soh, "Learning CNN-LSTM architectures for image caption generation," Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep, vol. 1, 2016.

[5] J. Bineeshia, "Image Caption Generation Using CNN-LSTM Based Approach," 2021.

[6] M. J. Panicker, V. Upadhayay, G. Sethi, and V. Mathur, "Image caption generator," Int. J. Innov. Technol. Explore. Eng., vol. 10, no. 3, 2021.

[7] X. Yin, C. Liu, and X. Fang, "Sentiment analysis based on BiGRU information enhancement," in Journal of Physics: Conference Series, 2021, vol. 1748, no. 3, p. 32054.

[8] X. Xu, H. Dinkel, M. Wu, and K. Yu, "A CRNN-GRU Based Reinforcement Learning Approach to Audio Captioning.," in DCASE, 2020, pp. 225–229.

[9] C. Wall, L. Zhang, Y. Yu, A. Kumar, and R. Gao, "A deep ensemble neural network with attention mechanisms for lung abnormality classification using audio inputs," Sensors, vol. 22, no. 15, p. 5566, 2022.

[10] X. Xu, H. Dinkel, M. Wu, and K. Yu, "The SJTU submission for DCASE2020 task 6: A CRNN-GRU based reinforcement learning approach to the audio caption," DCASE2020 Challenge, Tech. Rep., 2020.

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

[12] M. Wang, L. Song, X. Yang, and C. Luo, "A parallel-fusion RNN-LSTM architecture for image caption generation," in 2016 IEEE international conference on image processing (ICIP), 2016, pp. 4448–4452.

[13] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," IEEE Trans. Image Process., vol. 28, no. 1, pp. 32–44, 2018.

[14] Y. Lee, J. Kwon, Y. Lee, H. Park, H. Cho, and J. Park, "Deep learning in the medical domain: predicting cardiac arrest using deep learning," Acute Crit. care, vol. 33, no. 3, pp. 117–120, 2018.

[15] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in International conference on machine learning, 2015, pp. 2048–2057.

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

[17] C. Yan et al., "Task-adaptive attention for image captioning," IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 1, pp. 43–51, 2021.

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

[19] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086. International conference on computational intelligence in data science (ICCIDS), 2017, pp. 1–6.

[20] P. Mathur, A. Gill, A. Yadav, A. Mishra, and N. K. Bansode, "Camera2Caption: a real-time image caption generator," in 2017

[21] R. Subash, R. Jebakumar, Y. Kamdar, and N. Bhatt, "Automatic image captioning using convolution neural networks and LSTM," in Journal of Physics: Conference Series, 2019, vol. 1362, no. 1, p. 12096.

[22] G. Sharma, P. Kalena, N. Malde, A. Nair, and S. Parkar, "Visual image caption generator using deep learning," 2019.

[23] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille. Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090, 2014

[24] Amritkar, C., & Jabade, V. (2018, August). Image caption generation using deep learning technique. In 2018 fourth international conference on computing communication control and automation (ICCUBEA) (pp. 1-4). IEEE.

[25] K. Xu (2015) Show, attend and tell: Neural image caption generation with visual attention. inProc. Int. Conf. Mach. Learn.

[26] Andrej Karpathy, Li Fei Fei (2015) Deep Visual-Semantic Alignments for Generating Image Descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence (April 2017), vol 39, issue 4:664 – 676

AUTHORS

First Author – Fida Hussain Dahri, Research Scholar, Quest University Nawabshah, Sindh Pakistan

Second Author – Dr. Asghar Ali Chandio, Ph.D., Quest University Nawabshah, Sindh Pakistan

Third Author – Dr. Nisar Ahmed Dahri, Ph.D., Quest University Nawabshah, Sindh Pakistan

Fourth Author – Muhammad Ali Soomro, Ph.D., Quest University Nawabshah, Sindh Pakistan

Correspondence Author – Fida Hussain Dahri, Research Scholar, Quest University Nawabshah, Sindh Pakistan