Generating Human-Like Descriptions for the Given Image Using Deep Learning

Tanvi S. Laddha[1*], *Darshak* G. Thakore[2] and *Udesang* K. Jaliya[3]

[1]MTech Student, Dept of Computer Engg, BVM College, Vallabh Vidyanagar, Anand – 388120, Gujarat, India.
[2]Prof. & Head of Computer Dept, BVM College, Vallabh Vidyanagar, Anand – 388120, Gujarat, India.
[3]Assistant Prof., Dept of Computer Engg, BVM College, Vallabh Vidyanagar, Anand – 388120, Gujarat, India.

Abstract: One of the most prominent applications in the field of computer vision and natural language processing research is image captioner. The paper includes an exhaustive review of the literature on image captioning and the implementation using attention-based encoder-decoder model. The process of depicting an image with textual explanations is known as image captioning. The problem has seen extensive use of encoder-decoder frameworks. In this study, Deep Convolutional Neural Network (CNN) for image classification and Recurrent Neural Network (RNN) for sequence modeling are combined to build a single network that creates descriptions of images using the Microsoft Common Objects in Context Dataset (MSCOCO Dataset). Because of RNNs being computationally expensive to train and assess, memory is often restricted to a few items. By highlighting the most important components of an input image, the Attention model had been used to address this issue. The model was developed using Nvidia Quadro RTX5000 GPU (CUDA), which received the Bleu-1 score of 0.5793 for the 100 generated sentences. The captions generated by the model on the testing dataset labeled nearly all of the objects in the image and were sufficiently like the actual captions in the annotations, even on images outside of the testing dataset.

1 Introduction

The quantity of information that humans can digest in a single instant is enormous. Pictures, videos, and any other textual content are most likely included in this information. Humans utilize their natural language to convey images, and each image contains a wealth of information that may be interpreted and processed [1].

*Corresponding author: tladdha08@gmail.com

The formation of a relationship between the extracted objects is required for sentence formation. The feature extraction approach may be used to extract image information [2]. The possibility of streamlining several coherent jobs opens up if the same may be accomplished by computers. However, creating captions for images is a time-consuming and difficult process for the technology of today [3]. Implementing an encoder-decoder architecture is one of the most effective ways to caption images.

A language generation model, such as Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), or one of its variations, is used to decode the high-level representation of the pictures once they have been encoded. Many models, including ResNet and its variants, InceptionV3, and other VGG variants, have been introduced. Flickr datasets, MSCOCO datasets which contain a large number of refined photos in .jpg format, are the most often used datasets for image caption training [1]. The applications of image captioning are mentioned below [4]:

- The close captioning process for the creation, editing, distribution, and preservation of digital information is automated and sped up by the picture captioning paradigm.
- When picture descriptions are read aloud to visually challenged people, it helps them understand their surroundings better. This is made possible by an AI-powered image caption generator.
- The picture captioning methodology expedites the production of subtitles and frees executives to concentrate on other crucial activities.

2. Related Works

Among all the methods, a generative-based method and a retrieval method are the most popular approaches. The Im2Txt model [5], arguably one of the finest retrieval technique models, was created and suggested by Girish Kulkarni, Vicente Ordonez, and Tamara L Berg. Their approach mainly consisted of two components: image matching and caption creation. The model in the retrieval-based approach receives an input image, and as a result, the database, which includes the photos and the necessary descriptions, will be searched for matching images. Once the photos have been located, they are evaluated against matched images and high-level objects from the original input photographs [1]. The fundamental drawback of such a retrieval-based approach is that it can only provide captions that are currently present in the dataset and cannot produce really original captions for unseen context [6]. Generative-based models address the drawbacks of retrieval-based methods. It's used to come up with creative descriptions for the pictures. The generative-based models are either pipeline-based models or end-to-end based models.

Image Captioning using Deep Learning [7] by Yukti Sanjay Jain, Tanisha Dhopeshwar, Surpreet Kaur Chaddha and Vrushali Pagire suggested the work that employs a dense attention model with a CNN encoder and RNN decoder architecture, where the encoder is in charge of object recognition and feature extraction and the decoder is in charge of creating the following words for picture description. The original seq2seq model's inability to effectively interpret the output for lengthy and complex text inputs was a flaw that was addressed by the introduction of the attention model.

As per Show, Attend to everything, and tell: Image Captioning with More Thorough Image Understanding [8] by Zahra Karimpour, Amirm Sarfi, Nader Asadi, Fahimeh Ghasemian, the decoder was given the output of the convolutional network, which was then succeeded by a linear layer. First, they updated the encoder since the linear layer that succeeded the encoder in the prior iteration was suspected of narrowing up the information in the image. This time, they utilize the channels of the encoder network's last convolutional layer, which for generators offer little extra information [8], [9]. Moreover, the decoder would concentrate on a particular area of the picture at each stage of the process utilizing an attention mechanism depending on its current concealed state. They also demonstrated how significantly the RNN's performance would be improved by adding an image as an additional input.

In Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation [10], to minimize overfitting of the model, S. Katiyar and S. Borgohain suggested employing Image Data Augmentation in combination to the Inception-ResNet Convolutional Neural Network as encoder, Hierarchical Context based Word Embeddings for word representations, and a Deep Stacked Long Short-Term Memory network as decoder. In addition to perspective transformations on the photos for data augmentation, they also applied horizontal and vertical flipping and tested their suggested solutions using the Encoder-Decoder and Soft Attention image captioning frameworks.

2.1 Methodology

Large quantities of features collected from the original image are compressed by neural networks into a smaller and RNN-compatible feature vector. CNN is sometimes known as "Encoder" for this reason [4]. The following pre-trained models can be used to extract features:

2.1.1. VGG16

The Oxford Visual Geometry Group, or VGG, model, which won the 2014 ImageNet competition, which has 16 convolutional layers, including 13 convolutional layers, 2 fully connected layers, and 1 SoftMax activation layer, this model is known as the VGG16 model [12]. The VGG16 basic model is to be modified by eliminating the final layer in order to generate captions [13]. The primary drawback of vgg16 is the vanishing gradient problem, which Resnet50 remedied.

2.1.2. InceptionV3

The Inception-V3 model employs several kernels with sizes of 1*1, 3*3, and 5*5. Inception-v3 employs batch normalization in the auxiliary classes and accepts a fixed input of a 299*299 RGB picture. The channel is aggregated following the convolutional

operation, and the fusion operator is then applied to the output of the preceding layer. As a result, it aids in lowering overfitting and enhancing the network's flexibility [13].

2.1.3. ResNet

ResNet-101 is a 101-layered deep convolutional neural network. Due to the enormous number of layers, the network learns detailed feature representations for a variety of pictures. The input picture size for the network is 224*224 pixels. A 7*7*2048 feature vector is produced [14]. With ResNets, the gradients can pass right through the backward skip connections from earlier layers to first filters [15].

2.1.4. Attention Mechanism

A pre-trained Convolutional Neural Network (Encoder), used in "traditional" image captioning systems, would encode the picture and create a hidden state [16]. Some of the characteristics of the encoded image are represented in the concealed state. The Decoder cannot effectively generate distinct words for different areas of the image when Attention is absent because it treats every aspect of the image identically when creating the output word [17], [16]. The decoder's task is to condense all of the incoming into one, fixed vector. Here, it is important to consider whether a given vector can adequately capture all of the essential data associated with that image. The input and output sequences are aligned using the attention mechanism, with a feed-forward network parameterizing the alignment score [18]. The bottleneck performance of classic encoder-decoder systems was addressed with the introduction of the Bahdanau attention, which led to substantial advancements over the earlier method.

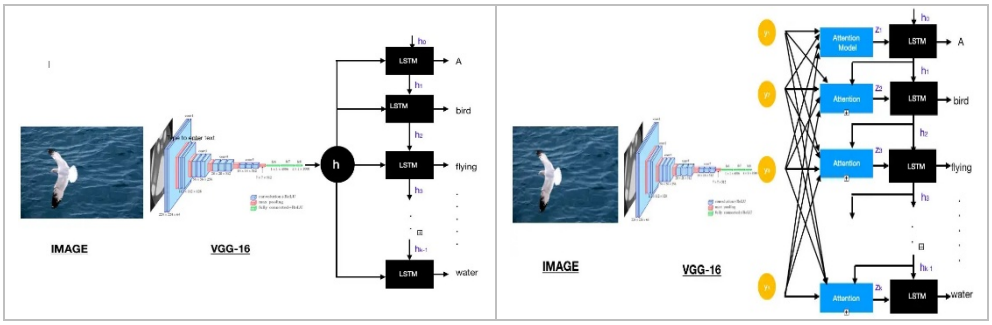


Fig. 1. Image captioning models (a) A "traditional" approach and (b) Attention Mechanism-based approach [26].

The basic figure of the "traditional" model for picture captioning is still recognizable, although there is a new layer of attention model. The fundamental idea behind the Bahdanau attention mechanism is to prioritize specific input vectors within an input sequence based on attention weights. The more important it is for a word to be produced at the following timestep, the higher its weight should be [13], [17].

2.2. Evaluation metrics

BLEU, abbreviation for The Bilingual Evaluation Understudy Score, is a measure for assessing a produced sentence to a reference sentence [22]. However, it is easy to calculate,

comprehend, and has a number of strong advantages. Even two individuals would probably come up with many phrase alternatives for a given issue and would seldom find a perfect fit [23]. To determine the number of matches, BLEU compares the n-grams of the candidate translation and the reference translation. These matches don't depend on the positions in which they take place [24]. The working of Bleu score can be well understood by the formula given by Shinde et al. [25].

$$modified\ ngram\ precision = \frac{number\ of\ times\ ngram\ occurs\ in\ reference}{total\ number\ of\ ngrams\ in\ hypothesis} \quad (1)$$

Hence, the sum of the n-gram counts for all the reference sentences in the corpus, divided by the number of n-grams in the hypothesis, is known as modified ngram precision [24].

3 Dataset Details

Flickr8k: 8091 photos of various sizes and shapes in the JPG format are included in Flickr8k, along with five distinct captions. 6000 of which are utilized for training, 1000 for testing, and 1000 for validation [19].

Flickr30k: Over 31K photos make up the dataset. To train the model, 29K photos are utilized, and 1000 images each are used for testing and verifying the model with five reference sentences [20].

MSCOCO: A sizable object detection, segmentation, key-point detection, and captioning dataset is the MS COCO (Microsoft Common Objects in Context) dataset. In 2014, the MSCOCO dataset's first version was made public. The splits from the 2014 version were adjusted later in the 2017 edition. There are 282K photos in the collection [21].

Table 1. Dataset Details

Dataset	Training set	Testing set	Validation set	Total
Flickr8k	6000	1000	1000	8091
Flickr30k	29000	1000	1000	31783
MSCOCO	236574	40670	5000	282244

4 Implementation Details and Results

The implementation incorporates the feature extraction from the pretrained models like VGG16, InceptionV3 and ResNet152 followed by the text preprocessing and text prediction to generate the captions for the images over the three different datasets mentioned above.

The flickr8k dataset, which contains 8k images and 5 captions for each, is utilized in the study to create captions for the provided images.

The VGG16 Neural network is used in restructured form to extract the features from the picture and the text captions for input. To anticipate the following word of the caption, the characteristics are then concatenated. For images, CNN is used, and for text, LSTM. The model was trained on kaggle GPU P100 for epochs = 50 and batch_size = 32. The Bleu score obtained after training the model was BLEU-1: 0.546441 & BLEU-2: 0.323487.

The InceptionV3 model was used to extract features and was trained with the same flickr8k dataset on kaggle GPU P100 for epochs = 50 and batch_size = 3.

The Beam Search and Greedy Search were used to predict the captions. When using Greedy Search, just the top word at each slot is selected. Beam Search, on the other hand, broadens this and selects the finest "N" terms [26].

The captions generated by the model are shown below in figure 2.

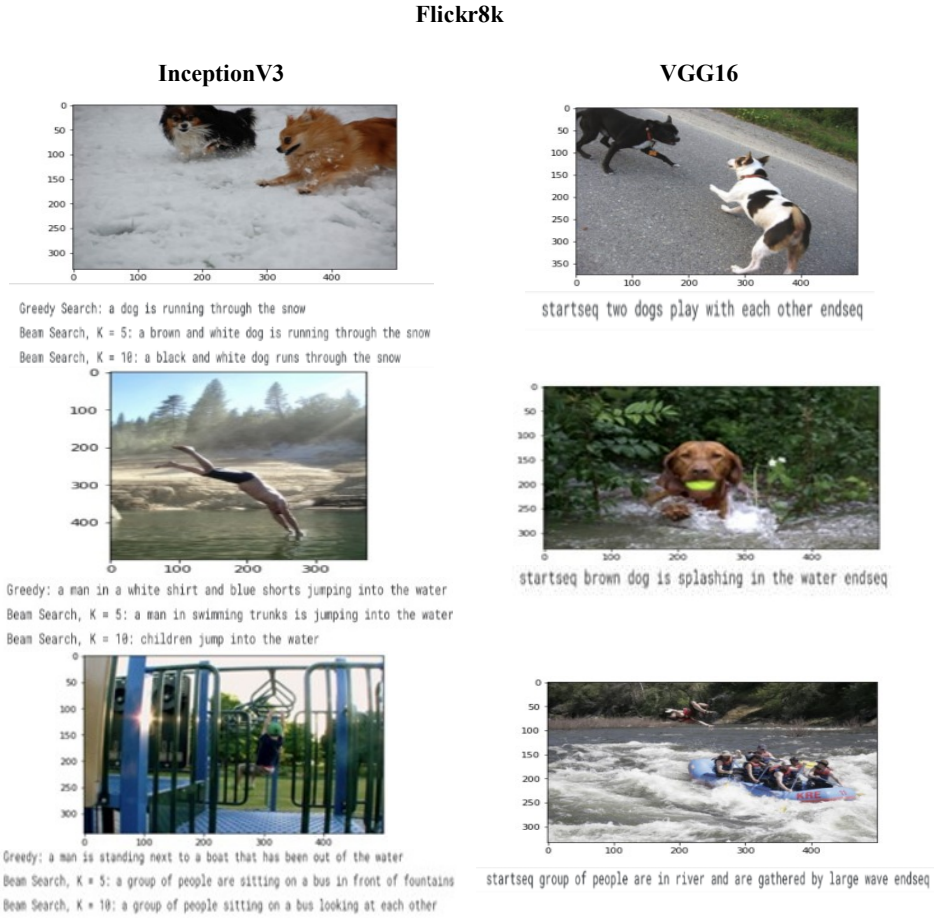


Fig. 2. The captions generated by the inceptionV3 and VGG16 model on the flickr8k dataset

The next was, flickr30k dataset, which contains 31k images and 5 captions for each. And again, the same VGG16 and InceptionV3 neural networks which are pre-trained on the

imagenet dataset by Google were used for feature extraction: VGG16 was trained on kaggle GPU P100 for epochs = 20 and batch_size = 32. The Bleu score yield was BLEU-1: 0.532794 & BLEU-2: 0.283582. The InceptionV3 was trained on kaggle GPU T4x2 for epochs = 30 and batch_size = 12. The Beam Search and Greedy Search were used to predict the captions. The captions generated by the model on flickr30k dataset are shown below in figure 3.

Flickr30k

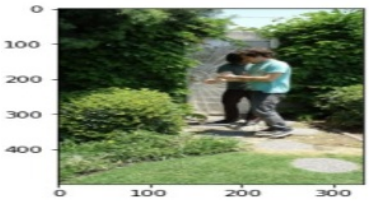
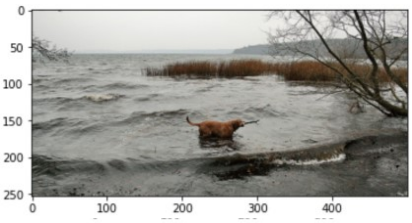
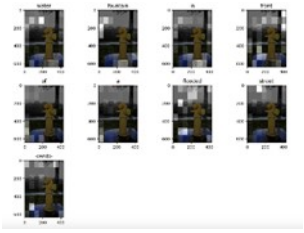
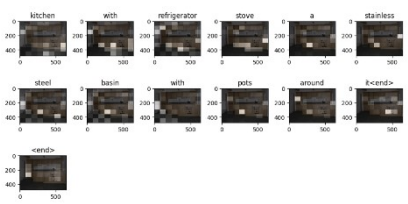
InceptionV3	**VGG16**
 group of people are walking down the street	 startseq man in blue shirt and jeans is sitting on the ground in front of tree endseq
 two greyhounds racing each other	 startseq man in white shirt is cooking food in restaurant kitchen endseq
 two women are walking down the street	 startseq man in black shirt and jeans is playing an electric guitar endseq
 dog is running through the water	 startseq man in blue shirt and black hat is sitting on the beach looking at the camera endseq

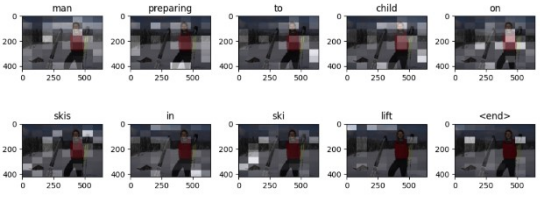
Fig. 3. The captions generated by the InceptionV3 and VGG16 model on the flickr30k dataset

To accurately and completely describe the major semantic content of an image and demonstrate the variations in content exhibited by other scene photos, image captioning is required to not only recognise items and deduce the relationships among them but also to take into account context information between the target objects, objects, and the image scene surroundings [27].

To simultaneously address the objectives of object detection, understanding the relationships between the discovered items, and image captioning, visual attention is used on MSCOCO dataset. The InceptionV3 model is trained over Nvidia RTX5000 GPU with accelerated computing (CUDA) for epochs = 10, batch_size = 64. This is because the feature vector extracted from InceptionV3 has a shape (64, 2048). The loss = 0.007818 (at the end of 10 epochs) where Time taken for 1 epoch 1579.94 sec







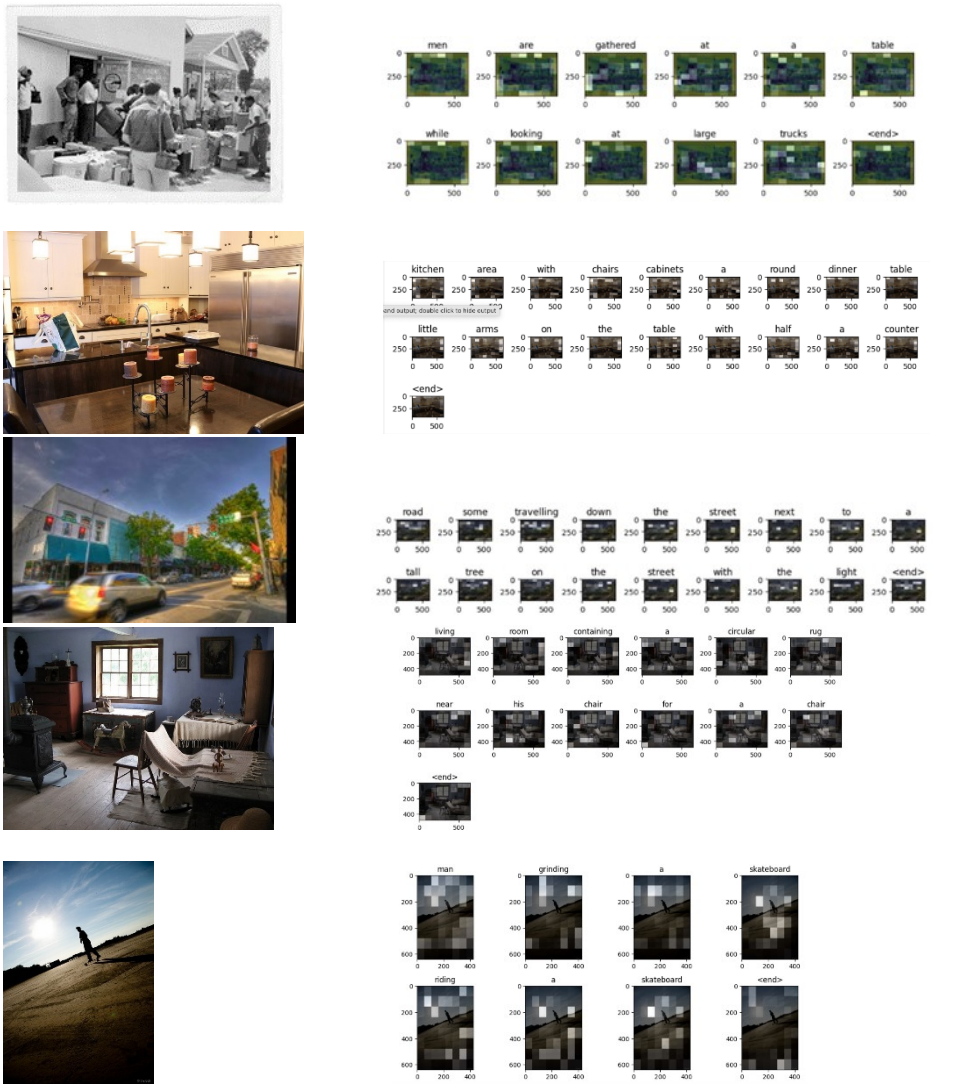


Fig. 4. Attention mechanism implemented on InceptionV3 model using MSCOCO dataset to identify the various objects in each image.

The method of pre-processing an image is rather simple. After putting the image into the training folder, it is processed using the transform function, which outlines how to prepare the images for use as input to the CNN encoder by converting them to PyTorch tensors. Additionally, the captions need to be prepared for training and pre-processed. Using the nltk.tokenize.word tokenize method, a list of string-valued tokens is obtained once all of the letters in the caption are changed to lowercase. The string-valued caption is transformed into an array of integers, which is then transformed into a PyTorch tensor and cast to long type.

Finally, the model Resnet152 is used for feature extraction and EncoderCNN and DecoderRNN are trained by adjusting the hyperparameters batch_size = 128 and num_of_epochs = 6. After 6 epochs the captions predicted are shown below as good, average and excellent based on the quality of the caption produced as per human evaluation. The Bleu score obtained after training the model on the Nvidia Quadro RTX5000 GPU is Bleu-1: 0.5793, Bleu-2: 0.4043, Bleu-3: 0.2785, Bleu-4: 0.1908, the CIDEr: 0.5997, ROUGE: 0.3962. It should be clear that the pictures used for testing and training the model must be semantically relevant. For instance, if the model was trained on pictures of cats, dogs,etc. it shouldn't be tested on pictures of airplanes, waterfalls, etc.

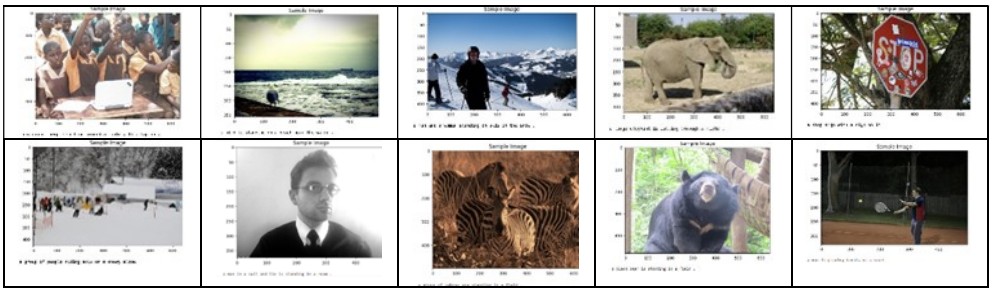


Fig. 5. The captions for the given images that fall into the "excellent" category, according to human review.

Good

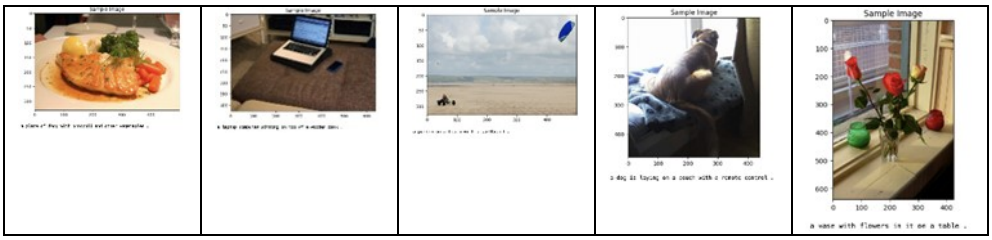


Fig. 6. The captions for the given images that fall into the "good" category, according to human review.

Average

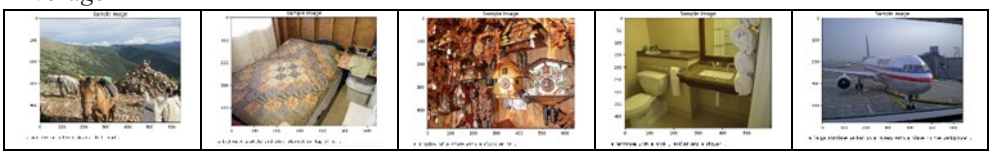


Fig. 7. The captions for the given images that fall into the "average" category, according to human review.

5 Challenges and Results

The same photograph might have several captions created by different people. One of the most significant objectives of computer vision is automatically creating captions for a picture, particularly called as, scene interpretation. Additionally, when identifying the distinct items in an image, caption generation models need to be strong enough to explain the connections between those objects in natural language. So that when the image having the same objects placed into different context or background is given, the model can easily

recognize them and accurately represent the scene in the photograph. The process of analyzing such areas and objects and determining what is truly happening in the image is called image comprehension [11]. In several cases, it tends to get biased towards the more dominant colors in the image, which led to certain words being incorrectly predicted as shown in figure 8 (a) and figure 8 (b). This issue is yet to be addressed.

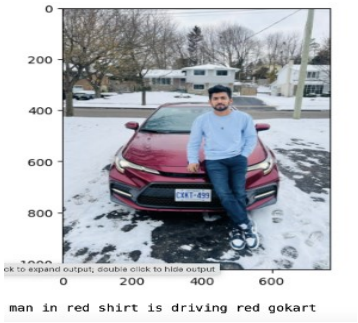


Fig. 8. (a) *shows the domination of red color in the image*

Fig. 8. (b) *shows the domination of black color in the image*

Table 2. Results

Dataset	Model	**Bleu-1**	**Bleu-2**	**Bleu-3**	**Bleu-4**	**CIDEr**	**ROUGE**
MSCOCO	ResNet152	0.5793	0.4043	0.2785	0.1908	0.5997	0.3962
Flickr8k	VGG16	0.5464	0.3234	-	-	-	-
Flickr30k	VGG16	0.5327	0.2835	-	-	-	-

6 Conclusion

Image captioning is the method of producing a textual description of an image utilizing both computer vision and natural language processing programs. It necessitates dealing with both text and pictures. As mentioned in the aforementioned study, Recurrent Neural Network (RNN) is used to decode feature and word representations and construct language models after Convolutional Neural Network (CNN) encodes pictures into latent space representations and the network's capacity to anticipate better captions is enhanced by attention. The ResNet152 model trained on MSCOCO dataset predicted the almost accurate captions. The Bahdanau's attention model also focused on the objects that were zeroed out by the classical model. Hence, the best bleu score obtained after evaluating the sentences generated by the implemented model was 0.579 over the 100 generated sentences.

References

1. J. Sudhakar, V. V. Iyer, and S. T. Sharmila, "Image Caption Generation using Deep Neural Networks," in *2022 International Conference for Advancement in Technology (ICONAT)*, Jan. 2022, pp. 1–3. doi: 10.1109/ICONAT53423.2022.9726074.
2. T. Patel, "Object Detection Based Automatic Image Captioning using Deep Learning," Comput. Eng.
3. Alex Krizhevsky, Alex Krizhevsky, Google Inc, View Profile, Alex Krizhevsky, and Alex Krizhevsky, "ImageNet classification with deep convolutional neural networks".
A. Team, "Building and Deploying an AI-powered Image Caption Generator," *AI Oodles*, Apr. 08, 2020. https://artificialintelligence.oodles.io/blogs/ai-powered-image-caption-generator/ (accessed Jan. 20, 2023).
4. "Image2Text | Proceedings of the 24th ACM international conference on Multimedia." https://dl.acm.org/doi/10.1145/2964284.2973831 (accessed Oct. 31, 2022).
5. "'Image Retrieval Using Image Captioning' by Nivetha Vijayaraju." https://scholarworks.sjsu.edu/etd_projects/687/ (accessed Oct. 31, 2022).
6. Y. S. Jain, T. Dhopeshwar, S. K. Chadha, and V. Pagire, "Image Captioning using Deep Learning," in *2021 International Conference on Computational Performance Evaluation (ComPE)*, Dec. 2021, pp. 040–044. doi: 10.1109/ComPE53109.2021.9751818.
7. Z. Karimpour, Amirm. Sarfi, N. Asadi, and F. Ghasemian, "Show, Attend to Everything, and Tell: Image Captioning with More Thorough Image Understanding," in *2020 10th International Conference on Computer and Knowledge Engineering (ICCKE)*, Oct. 2020, pp. 001–005. doi: 10.1109/ICCKE50421.2020.9303609.
8. "(PDF) Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." https://www.researchgate.net/publication/272194766_Show_Attend_and_Tell_Neural_Image_Caption_Generation_with_Visual_Attention (accessed Oct. 31, 2022).
9. S. Katiyar and S. Borgohain, *Image Captioning using Deep Stacked LSTMs, Contextual Word Embeddings and Data Augmentation*. 2021.
10. "BMorse-BYU-iu-active-contours.pdf." Accessed: Jan. 18, 2023. [Online]. Available: https://www.sci.utah.edu/~gerig/CS6640-F2012/Materials/BMorse-BYU-iu-active-contours.pdf
11. T. V. Sneha and D. S. J. Rani, "LSTM-VGG-16: A Novel and Modular Model for Image Captioning Using Deep Learning Approaches," vol. 12, no. 11.
12. S. Ayoub, Y. Gulzar, F. A. Reegu, and S. Turaev, "Generating Image Captions Using Bahdanau Attention Mechanism and Transfer Learning," *Symmetry*, vol. 14, no. 12, Art. no. 12, Dec. 2022, doi: 10.3390/sym14122681.
13. R. Khan, M. S. Islam, K. Kanwal, M. Iqbal, Md. I. Hossain, and Z. Ye, "A Deep Neural Framework for Image Caption Generation Using GRU-Based Attention Mechanism," 2022, doi: 10.48550/ARXIV.2203.01594.
14. P. Ruiz, "Understanding and visualizing ResNets," *Medium*, Apr. 23, 2019. https://towardsdatascience.com/understanding-and-visualizing-resnets-442284831be8 (accessed Jan. 24, 2023).
15. K. Doshi, "Image Captions with Attention in Tensorflow, Step-by-step," *Medium*, Apr. 30, 2021. https://towardsdatascience.com/image-captions-with-attention-in-tensorflow-step-by-step-927dad3569fa (accessed Oct. 06, 2022).

16. S. Sarkar, "Image Captioning using Attention Mechanism," *The Startup*, Jun. 15, 2021. https://medium.com/swlh/image-captioning-using-attention-mechanism-f3d7fc96eb0e (accessed Jan. 25, 2023).
17. T. Gautam, "Attention Mechanism For Image Caption Generation in Python," *Analytics Vidhya*, Nov. 20, 2020. https://www.analyticsvidhya.com/blog/2020/11/attention-mechanism-for-caption-generation/ (accessed Feb. 01, 2023).
18. "Flickr 8k Dataset." https://www.kaggle.com/datasets/adityajn105/flickr8k (accessed Feb. 01, 2023).
19. "Flickr30k Dataset," *Machine Learning Datasets*. https://datasets.activeloop.ai/docs/ml/datasets/flickr30k-dataset/ (accessed Feb. 01, 2023).
20. "Papers with Code - COCO Dataset." https://paperswithcode.com/dataset/coco (accessed Feb. 01, 2023).
21. J. Brownlee, "A Gentle Introduction to Calculating the BLEU Score for Text in Python," *MachineLearningMastery.com*, Nov. 19, 2017. https://machinelearningmastery.com/calculate-bleu-score-for-text-python/ (accessed Feb. 02, 2023).
22. "Foundations of NLP Explained — Bleu Score and WER Metrics | by Ketan Doshi | Towards Data Science." https://towardsdatascience.com/foundations-of-nlp-explained-bleu-score-and-wer-metrics-1a5ba06d812b (accessed Feb. 02, 2023).
23. R. Khandelwal, "BLEU — Bilingual Evaluation Understudy," *Medium*, Jan. 26, 2020. https://towardsdatascience.com/bleu-bilingual-evaluation-understudy-2b4eab9bcfd1 (accessed Feb. 02, 2023).
24. "Image Captioning With Flickr8k Dataset & BLEU | by Raman Shinde | Medium." https://medium.com/@raman.shinde15/image-captioning-with-flickr8k-dataset-bleu-4bcba0b52926 (accessed Feb. 03, 2023)
25. K. Doshi, "Foundations of NLP Explained Visually: Beam Search, How it Works," *Medium*, May 21, 2021. https://towardsdatascience.com/foundations-of-nlp-explained-visually-beam-search-how-it-works-1586b9849a24 (accessed Feb. 01, 2023).
26. P. Tian, H. Mo, and L. Jiang, "Image Caption Generation Using Multi-Level Semantic Context Information," *Symmetry*, vol. 13, no. 7, Art. no. 7, Jul. 2021, doi: 10.3390/sym13071184.