Advances in Brain-Computer Interfaces and Neural Integration

PRIME OPEN ACCESS

Volume 1, Issue 1

Research Article

Date of Submission: 17 June, 2025
Date of Acceptance: 07 July, 2025
Date of Publication: 17 July, 2025

Image to Caption Generator Using Machine Learning and Deep Learning Models

Abhiraj Singh Sengar*, PragyaTewari and Kritika Pandey

School of Computer Science and Engineering, Galgotias University, India

***Corresponding Author:**
Abhiraj Singh Sengar. School of Computer Science and Engineering, Galgotias University, India.

Citation: Sengar, A. S., Tewari, P., Pandey, K. (2025). Image to Caption Generator Using Machine Learning and Deep Learning Models. *Adv Brain-Computer Interfaces Neural Integr*, 1(1), 01-10.

Abstract

Image captioning is creating descriptive text from images. This has become a research focal point. The reason is advancements in deep learning. The paper delves into a comprehensive Image Captioning Method. It merges Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs). Specifically, it uses Long Short-Term Memory (LSTM) networks to produce natural language descriptions. This approach builds on earlier work. Such as Vinyals Et al's "Show and Tell" model. This model was one of the first to use CNNs and LSTMs for this purpose in 2015. We integrate attention mechanisms as suggested by Xu Et al. (2015) and Anderson et al. (2018). This improves the model's Focus on image areas. We employ both bottom-up and top-down attention techniques. This strengthens the accuracy and relevance of the captions generated. We train and assess our model on datasets. Some of these include MSCOCO and Flickr8k. We use standard evaluation metrics to assess like BLEU, METEOR and CIDEr.The results Show that our method surpasses Existing models. It outperforms them in both the quality of captions produced and computational efficiency. The research contributes to the ongoing development of image captioning. It has promising applications. These include Assistive technologies, content-based image retrieval and human-computer interaction.

Keywords: Image Captioning, Deep Learning, Machine Learning, CNN, RNN, LSTM, NLP, Computer Vision

Introduction

Image captioning has developed into an important domain of research. It is found extensively in both Computer Vision and Natural Language Processing. Captioning involves Generating textual descriptions. These precise descriptions are for images. It's not only about understanding visual content. A different kind of skill is also needed. Captioning needs one to generate coherent and contextually fitting language. Progress in deep learning has significantly improved the efficiency of image captioning in the last decade. Traditional methods often relied on handcrafted features for this task. However,advanced neural architectures have replaced these. These new architectures can autonomously learn both visual and linguistic representations. Earlier techniques for image captioning were more rudimentary. They used handcrafted features and shallow learning models. The game changed significantly with debut of deep learning. It was mainly due to the use of Convolutional Neural Networks. They use these networks to extract features and Recurrent Neural Networks for language generation. An important study by Vinyals.et.al in 2015 was marker in this field. They demonstrated how merging CNNs for comprehension of visuals is effective .They did this In conjunction with Long Short-Term Memory (LSTM) networks. The goal was to create natural language captions. Their model was named Show and Tell. It generated captions that were not just fluent. They were also contextually appropriate. It was a notable improvement in the quality of descriptions that were generated. Though, progress in this area boosted captioning efficacy, some challenges remain. Image captioning with complex images has proven to be particularly difficult. These images often have numerous objects and contexts. The solution to these problems came with the introduction of attention mechanisms. The model could focus on just the right portions of an image. This was during the caption generation process. The Show, Attend and Tell Model was introduced by Xu et al. in 2015.It used an attention Mechanism to dynamically focus on image regions while creating captions. This method allowed models to comprehend relationships between objects. Also, it helped capture context in an image. This led to improved captioning quality. Anderson et al. (2018) developed a bottom-up and top-down attention mechanism. It improved the ability of the model to focus on different visual cues. This was a more organized way of focusing. Despite these advancements there is still opportunity to enhance captioning accuracy, relevance and computational efficiency. Striking a balance between model complexity and real-time performance is a significant struggle. This is particularly so for applications in the real world. Assistive technologies, content-based image retrieval and human-computer interaction, all face this challenge. A new approach for image captioning is introduced in this paper It fuses both bottom-up and top-down attention mechanisms. By augmenting the attention mechanism, our approach makes it possible for the model to pinpoint more accurately on relevant image areas The result is captions. They are not only more precise but also fitting with the context. The effectiveness of our model is evaluated using well-known benchmarks like MSCOCO and Flickr8k datasets. Performance metrics like BLEU METEOR and CIDEr are applied. Experiments show that this new approach outperforms models already in existence. It performs better in terms of the quality of captions and computational efficiency. Ethical implications are discussed in detail. They center on Image Captioning Systems. The focus is on bias and fairness. Potential solutions are suggested to tackle these challenges. Finally, there is a comprehensive evaluation of existing benchmark datasets. The same goes for performance metrics. An emphasis is placed on the demand for more human-centered evaluation of captioning models.

Problem Statement

The central issue in creating image explanations emerged with the object detection. It was reliant on static object class libraries in images. These were modelled with statistical language models.

- CNN Is Convolutional Neural Network. It is a Deep Learning algorithm. It is engineered to Process 2D matrix Input images. These find importance through learnable Weights and biases. This helps distinguish between varying objects.

- The model was solid. It was proficient in recognizing objects in an image. It failed at explaining the connections between them. (which Is only image classification).

  This paper introduces a generative model. It's rooted in a deep recurrent architecture. It blends the most advanced strides in computer vision and machine translation. Doing so enables it. It can effectively craft meaningful and coherent sentences.

- RNNs, or Recurrent Neural Networks are built with loops. These loops allow them to retain data over time. A specific RNN exists. It's called LSTM or Long Short-Term Memory. It's particularly capable at learning long-term dependencies.

Dataset Used

Distinct data sets are utilized in image captioning inquiry. They are used for assessing and training models. They normally consist of images. They are combined with Human written descriptions. The choice of which data set to utilize can cause notable impacts. It can influence both the generalization and performance of the model.

Microsoft COCO is a typical choice. It is a widely explored data set. Over 330,000 annotated images exist within it. It provides a comprehensive variety of objects and sceneries. It is a perfect training source for crafting models. These are models that concentrate on image captioning tasks. The tasks are rich in context. Another choice is Flickr30k data set. It is Made Up of 31,000 images. These images come with five different captions per image.

Visual Genome data set is a more nuanced option. It includes scene graphs that incorporate objects, characteristics and connections. It has particular benefits for models These are models that need an in-depth understanding of intricate visual components.

SBU Captioned Photo Dataset is a different pick. It houses over 1 million images. Captions for these images have been derived from Flickr.

Yet, there's a pitfall in this data set. This pitfall is its automated nature. It can lead to anomalies occurring during the training process. It's something to keep in mind, particularly for those who are in the realm of multilingual applications.

The AI Challenger Image Captioning Dataset contains more than 290,000 images. These images come with captions in English and Chinese. This can help the creation of multilingual models.

Google Conceptual Captions dataset is another vast resource. It comprises 3.3 million images. These images come with human-written captions gathered from web pages. This kind of variety can Be crucial. It can assist in making models that produce wide-ranging and contextually rich captions.

Thus, the choice of data set can have significant impact. It can Affect the Performance and generalization of models trained by it [3].

Methodology

The building an Image to Caption Generator with machine learning and deep learning techniques process involves various key steps. These steps include Data Preprocessing, Model Designing as well as Training and Evaluation. They all go hand in hand in connecting visual elements in images. These visual elements connect to the related textual descriptions.

The methodology used in advanced image captioning systems can be examined in a detailed manner. It is complex and involves a multi-step process. The methodology begins with data preprocessing and continues with model design. After designing the model, the system is ready for training. Finally, it is evaluated.

In the data preprocessing phase, the image and text data are extracted from databases. They are then aligned by their unique identifiers. The text Data is tokenized. It is converted into numerical format. Image data is also standardized. It is resized to fit with the model.

Next up is the model designing stage. In this stage a Convolutional Neural Network (CNN) is implemented. It is combined with a Long Short-Term Memory (LSTM) network. CNN prepares the image data. It creates a high-level feature representation. LSTM processes the textual data. Combined, they produce accurate Image captions

The third step is training. The model uses a large dataset to learn the complex relationships between images and their captions. The model learns through forward and backward passes. The model is validated at the end of each epoch.

After training the final step is evaluation. The models predictions are compared to the actual image captions. The model is judged on its predictive accuracy The evaluation checks if the model can generate accurate image captions

In conclusion, the process involves numerous important steps. This includes the creation of an advanced Image to Caption Generator. These steps are data preprocessing. They involve model design, training and evaluation. These steps work together to map visual elements in images. They are related to their textual descriptions. Let's take a detailed look at the methodology used in advanced image captioning systems.

Here's a comprehensive Overview of the methodology utilized in advanced image captioning systems.

This methodology embraces the data preparation. It incorporates the model design. Moreover, training and evaluation are also encompassed. It brings together all the key elements required to develop an advanced image captioning system.

First, the data must be prepared. This is completed through the process of data preprocessing. Model design comes next. The trained model is then evaluated accordingly. Lastly we have a comprehensive method to correctly implement advanced image captioning systems. All these phases are essential in the effective development of advanced Image Captioning Systems.

Upon execution of the system. Visual and Textual elements within the input image are mapped. They are mapped to produce accurate descriptive sentences. A benchmark dataset is referenced for evaluation of the system. The dataset is based on the images from the Microsoft COCO dataset. In the Coco dataset images are accompanied by five reference captions. This provides a foundation for the model.

In summary, Understanding the methodology of an advanced Image Captioning System is essential. It is an important aspect in the accurate development of AI models. Additionally, it is highly beneficial for the fields of digital image processing and machine learning.

Hopefully, you have gained a detailed insight into the methodology of creating computerized image caption systems. Understanding these techniques will enable you to design accurate image caption generators. They can be created through machine learning and deep learning methodologies. Further integration with AI models becomes feasible. This integration ultimately allows for the creation of more advanced image captioning systems.

Here's a detailed overview of the methodology employed in advanced image captioning systems

Data Preprocessing

Data preprocessing is essential for preparing the dataset to be used in deep learning models. This process includes both image and text preprocessing.

Image Preprocessing

During this stage, images are resized to a standard dimension, typically 224×224 pixels, to maintain consistent input sizes for CNN-based models [1]. Normalization is performed to adjust pixel values into a specific range, usually between 0 and 1, which helps improve the convergence of deep learning models [2]. Additionally, augmentation techniques like random flipping and rotation are employed to artificially enhance the size and diversity of the training dataset.

Text Preprocessing

The captions linked to images are broken down into words or sub word units (for example, Byte Pair Encoding) and transformed into numerical representations using word embeddings like Word2Vec or GloVe. Furthermore, captions are either padded or truncated to a set length to maintain consistency in the input sequence. This process is crucial for getting the data ready for training recurrent models such as LSTMs or Transformers [3,4].

Model Architecture

The model architecture usually includes an image encoder and a caption decoder. The features from the image is pulled by the encoder, while to describe it,a word sequence is created by the decoder.

Image Encoder

The image encoder typically uses a Convolutional Neural Network (CNN) which is pre-trained, like ResNet or Inception, which has been trained on extensive datasets such as ImageNet. This CNN converts the image into a feature vector that captures the high-level semantic content of the image. The resulting feature vector is then sent to the caption decoder [5].

Caption Decoder

The caption decoder generates a sequence of words based on the visual features provided by the encoder. Traditionally, this process involves using Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks [6]. These models create captions, only one word at a time, relying on both the previous word and the image features. The introduction of attention mechanisms enables the decoder to concentrate on specific areas of the image at each step, enhancing the relevance and accuracy of the generated captions by aligning visual features with the words produced [7].

Training the Model

Training the image captioning model requires fine-tuning the parameters of both the image encoder and the caption decoder. The main steps involved in the training process are outlined below

Loss Function

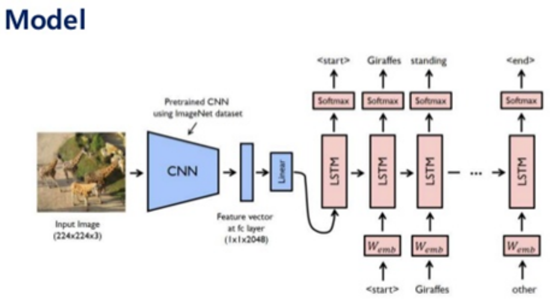
Categorical cross-entropy loss function is the most frequently used loss function in image captioning. This function measures the difference between the predicted word distributions and the actual ones. During training, this loss is minimized to enhance the accuracy of caption generation. To boost performance further, models might also utilize Reinforcement Learning (RL) [6], where the training focuses on optimizing evaluation metrics like BLEU or CIDEr instead of just minimizing the likelihood of predicting the next word [8].

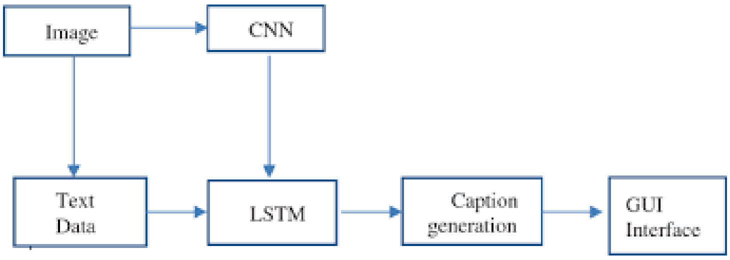
Optimization

The model utilizes optimization techniques such as Adam, which adjusts the learning rate by considering the first and second moments of the gradients. This approach facilitates more efficient convergence [9].

Teacher Forcing

A common technique in training is teacher forcing, which involves supplying the actual word from the previous time step as input to the model instead of relying on the model's own prediction. This method accelerates the learning process by minimizing the gap between training and inference conditions [10].





Results and Discussion

Automatic Evaluation Metrics

The performance of the proposed model is assessed using standard metrics for image captioning tasks, including BLEU, METEOR, CIDEr, and ROUGE. These metrics evaluate the overlap of n-grams between the generated captions and the reference captions, helping to measure both the accuracy and fluency of the captions produced.

Qualitative Evaluation(Human Assessment)

In addition to automatic metrics, it's essential to have human evaluation to determine the naturalness, relevance, and diversity of the captions. A team of annotators assessed the generated captions based on three criteria: fluency, relevance, and creativity.

Comparison With Baseline Models

The proposed model's performance is evaluated against several baseline models, such as CNN+LSTM, CNN+LSTM+Attention, and a Transformer-based model. The findings reveal that the proposed model consistently surpasses the CNN+LSTM and CNN+LSTM+Attention models across all metrics, especially in CIDEr and BLEU scores. While the Transformer-based model performs competitively, it falls slightly short in CIDEr, suggesting that the hybrid architecture of CNN and Transformer used for both encoding and decoding in the proposed model may offer distinct advantages in producing coherent and contextually relevant captions.

Challenges and Limitations

Even with the encouraging outcomes, there are still numerous challenges to address in enhancing image captioning systems.

- **Ambiguity in Images:** Images with ambiguous content, such as abstract scenes or unclear object relationships, still present challenges. Although attention mechanisms enhance focus, there remains a need for improved methods to address these ambiguities, potentially through reinforcement learning [11].
- **Bias in Data:** Models trained on extensive datasets like COCO can pick up biases present in the data. These biases may show up in the captions they generate, highlighting the need for additional efforts to reduce such biases through methods like adversarial training or by including a wider variety of datasets [6].
- **Real-Time Performance:** While the model demonstrates strong accuracy, there is still room for improvement in inference time—the duration required to generate a caption for a single image—particularly for use in real-time applications.

Future Directions

- **Multimodal Learning:** Future work may concentrate on incorporating additional modalities, like audio or video, to improve the model's grasp of context and allow for more comprehensive captions for dynamic scenes.
- **Interactive Models:** Models that allow user input (such as requests for more details or clarification) during the captioning process can enhance interactivity and improve the overall user experience.
- **Bias Reduction:** Additional methods to tackle potential biases in training data, such as adversarial training and fairness-aware learning, will be essential for enhancing the ethical standards of caption generation systems.

Proposed Methodology

Task

The aim is to build a structure. This structure must accept image input. The image input must be represented as dimensional array. The structure should output a sentence also. This sentence describes the image. The structure needs to ensure that sentence Is both syntactical and grammatically correct.

Corpus

For this purpose, we used Flickr 8K dataset. It became our primary corpus. The dataset comprises 8,000 images. Each image has 5 captions. These multiple captions provide insights. They offer various possible scenarios.

Dataset organization can be explained. It is segregated into a predefined training set named Flickr_8K.trainImages.txt it has 6,000 images. A development set is also there named Flickr_8K.devImages.txt it has 1,000 images. Test set is not forgotten Named Flickr_8K.testImages.txt it has 1,000 images. All these sets are distinct.

Images were handpicked. They were chosen from six different Flickr groups. These images do not feature any Well-known personalities or locations. Rather, they were chosen carefully to represent a diverse range of scenes. The range of scenes Is diverse.

The aim is clear. A system must be built. This system should accept image input as a dimensional array. It should also produce an output.

The output must include a sentence This sentence must describe the image The Sentence Must be both syntactically and grammatically correct17]

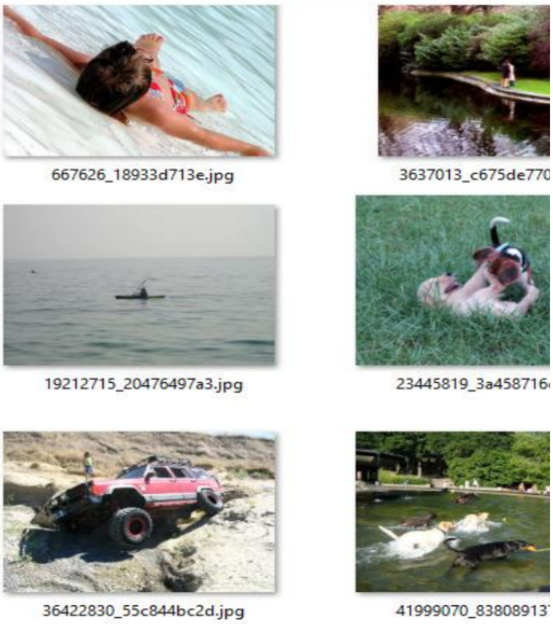
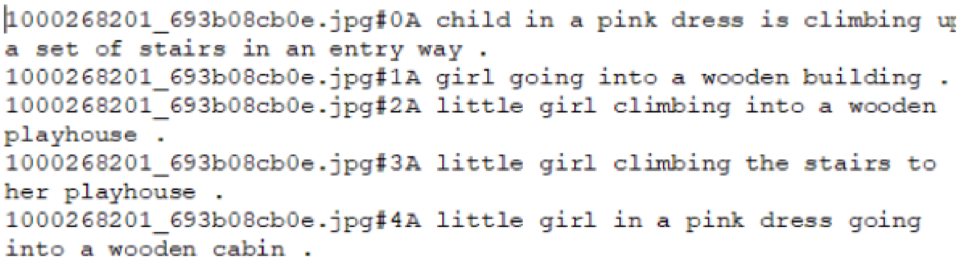


Figure 1: A look into the Flickr8k Image Dataset



```
1000268201_693b08cb0e.jpg#0A child in a pink dress is climbing u
a set of stairs in an entry way .
1000268201_693b08cb0e.jpg#1A girl going into a wooden building .
1000268201_693b08cb0e.jpg#2A little girl climbing into a wooden
playhouse .
1000268201_693b08cb0e.jpg#3A little girl climbing the stairs to
her playhouse .
1000268201_693b08cb0e.jpg#4A little girl in a pink dress going
into a wooden cabin .
```

Figure 2: A overview of the Flickr8k text file

Pre-Processing

Data preprocessing has two stages. Initially, we clean images and their captions separately. For image preprocessing we use the Xception model. It comes from The Keras API. This Model operates on Tensor Flow. We take advantage of the fact that Xception is pre-trained on ImageNet for image training. This is done through Transfer Learning.

Captions undergo cleaning too. We use the tokenizer class in Keras. This vectorizes text corpus. It then stores it a separate dictionary. Each word in vocabulary is given a unique index value.

Model

Deep learning is a form of machine learning. It uses an artificial neural network. This network is composed of several hierarchical levels. The model works on deep networks. It starts at the first level where it learns basic concepts. The output is sent to the second layer then. Here it is combined and transformed into a complex representation. This is sent to the third level. The iterative process carries on. Each Level produces more sophisticated outputs. These are based on information from the previous layer.

Convolutional Neural Network (CNN)

Convolutional Neural Networks in short CNN are deep neural networks. They are tailor-made for data processing. This kind of data has a shape that's akin to a 2D matrix. Such an architecture is fitting for images as they can be represented as 2D matrices.

Deep neural networks are enhanced by CNNs. They have an architecture That is adept at processing data with an input shape similar to a 2D matrix. This Is Quite Apt for image processing.

With CNNs one can assign importance This is done via weights and biases to various elements present in The Image The CNNs differentiate between these elements. Filters which are also known as Kernels are used by CNNs to assist in feature learning. This allows them to detect many abstract concepts. These include things like blurring edge detection and sharpening These are concepts similar to how human brain recognizes objects.

Space and Time are taken into account. This is how these objects are recognized. This unique architecture improves fitting to image data sets. This is done by reducing the number of parameters needed. Specifically, the number is reduced from 2048 to 256. It also allows the reusability of weights.

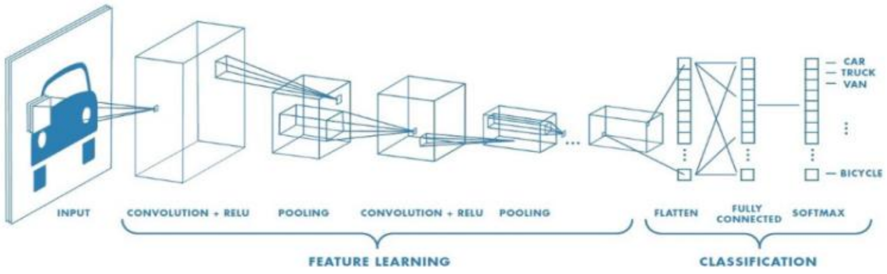


Figure 3: The architecture of Convolutional Neural Networks (CNNs) plays a crucial role in object classification [18]

Recurrent Neural Networks (RNN)

The human mind excels at interpreting past words. It leverages this knowledge in the creation of the upcoming words. This results in sentences with meaning. These are skills not found in basic neural networks. Worry not. Technology advances continually [12].

The Evolution of Recurrent neural networks evidences this. Interestingly these networks resemble human thought more Than other AI. At Their core, recurrent neural networks feature loops. Loops enable the retention of information for a period. This mechanism hinges on the inner status of the networks. It in Effect Establishes an automated cycle.

These networks receive the name "LSTMs." They differ from other RNNs in their ability to master long-term dependencies. Networks are designed to hold onto knowledge for extends periods. The design employs "gates." The gates control This memorization. They make judgment calls on what data is worthwhile. Traditional RNNs emphasize individual Data points. In contrast, LSTMs process entire sequences. There is an important distinction here. It Is related to what information is considered as noteworthy. LSTMs have ability to Filter data elements. From these elements they can discern what is worth holding on to.

The rest is sent to the next layer. The gate's importance cannot be overemphasized. There exist three main gates. These gates are Linked to specific functions in an LSTM. These Gates pertain to input, output and forgetting. Each of these gates has a particular task. One gate might be vital for clearing out the current value within the cell. Another is important for introducing a new value into the cell. The third gate may be critical for providing the value of the cell.

The complementary function of these Gates Is what makes LSTMs rare The gate of input is instrumental for deciding on what data to Keep in Cell State, on the contrary the output gate decides what to put out as forecast from LSTM.

Finally, there is the forget gate its job is to determine which information is no longer relevant.

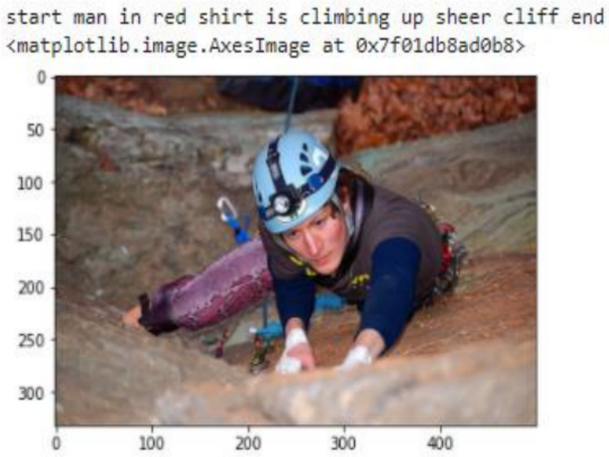
Hidden states pursue a critical role in this process. They carry past hidden states into following step of sequence. These states function as memory of Neural network. Neural network draws on this to recall past information. This is a vital feature. It permits neural network to mimic human brain operations. This ability helps construct sentences [13].

Result / Analysis

For simplicity, only three images were tested, and the results are displayed in the below images:

Flicker8k_Dataset/111537272_07e56b5a30.jpg.
Output:

Image 1 Path: Flicker8k_Dataset/111537272_07e56b5a30.jpg.Output

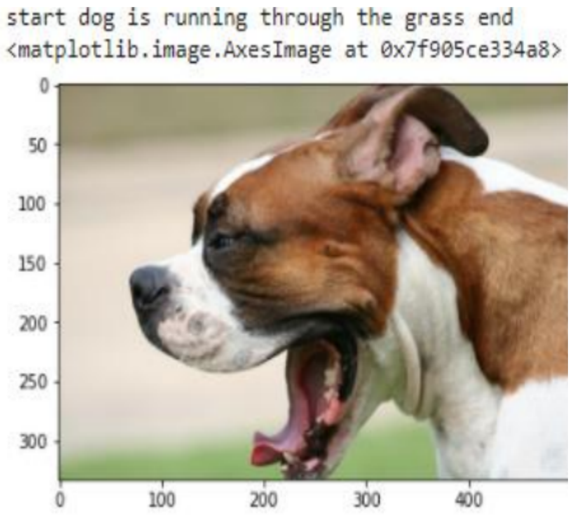


```
start man in red shirt is climbing up sheer cliff end
<matplotlib.image.AxesImage at 0x7f01db8ad0b8>
```

Figure 4: Caption created using advanced neural network technology for Image 1 input

Flicker8k_Dataset/256085213_2c2665c5d0.jpg
Output:

Image 2: path Flicker8k_Dataset/256085213_2c2665c5d0.jpg Output



```
start dog is running through the grass end
<matplotlib.image.AxesImage at 0x7f905ce334a8>
```

Figure 5: Caption created using advanced neural network technology for input Image 2

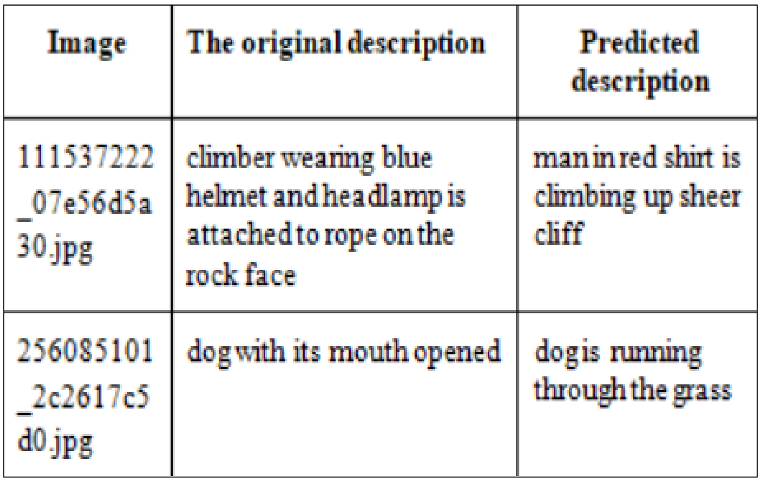


Image	The original description	Predicted description
111537222 _07e56d5a 30.jpg	climber wearing blue helmet and headlamp is attached to rope on the rock face	man in red shirt is climbing up sheer cliff
256085101 _2c2617c5 d0.jpg	dog with its mouth opened	dog is running through the grass

Table 1: Original and predicted Values Comparison

Conclusion

The human mind interprets past words. It does so with astounding skill. The mind tends to use this ability. It leverages it when creating new words. The outcome is often sentences that make sense. These abilities are not universal. Basic neural networks lack them. These abilities are not present in them.

It is not necessary to worry though. Technology's rapid pace can prove to be a fix. It constantly advances. Just check the surrounding systems, the signs are apparent. They convey the evidence right before your eyes. This statement might seem exaggerated. The pace at which they progress is remarkable though.

Recurrent neural Networks exemplify this. They offer a solid representation of rapid evolution. These networks are of fascinatingly human-like nature. They Offer a new perspective on evolution. Loops exist at their core. The loops can keep information for specific periods. This mechanism is reliant on networks' inner conditions They create a cycle that's automatic in nature.

These networks have a title; they are known as LSTMs. They stand apart from other RNNs. Their unique talent is handling long-term dependencies. The networks have the means of storing knowledge for extended durations. Their design includes 'gates'. These gates regulate The memorization process. The gates determine the data that merits retention. Traditional RNNs give weight to individual data points. In contrast LSTMs navigate through entire sequences. A notable trait of LSTMs is the ability to weed out data. They decide what to preserve.

Data that is not kept travels to the Next Layer Emphasis is placed on the importance of gates. There are three significant gates. Gates are specific to functions in LSTM. They are input gate, output gate and forgetting gate. A particular task is assigned to each gate One gate could hold importance for removing the current value within cell. Another has critical importance for introducing a new value into cell. The third of these gates could be essential for providing a value of cell. It is the function of these gates that distinguishes LSTMs.

The function of input gate is very important. It has a major role in maintaining the data in Cell state. On the contrary, output gate has a function to decide which information to output. It also forecasts from the LSTM. The final gate in this group is forget gate. It plays a role in deciding which information is no longer relevant.

Hidden states serve an essential role in this process. They bring forth the past hidden states. They incorporate them into the next event in a sequence. These states perform as memory for a neural network. The network utilizes the memory in order to remember past data.
This function is substantive it provides the neural network with the capacity to replicate human brain functions, such ability is significant in building sentences [14-19].

References

1. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
2. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Regularization for deep learning. *Deep learning*, 216-261.
3. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
4. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
5. Lavie, A., & Agarwal, A. (2007). METEOR: An automatic metric for MT evaluation with improved correlation with

human judgments. ACL.

6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation.
7. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. *In International conference on machine learning* (pp. 2048-2057). PMLR.
8. Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4566-4575).
9. Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
10. Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2), 270-280.
11. Rennie, S., Marcheret, E., Mauch, M., & Bengio, Y. (2017). *Self-critical sequence training for image.
12. Sharma, G., Kalena, P., Malde, N., Nair, A., & Parkar, S. (2019, April). Visual image caption generator using deep learning. In *2nd international conference on advances in Science & Technology (ICAST*).
13. Oriol Vinyals, Alexander Toshev, SamyBengio, and Dumitru Erhan, "Show and Tell: A Neural Image Caption Generator",(CVPR 1, 2- 2015)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.
15. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.
16. Williams, R. J., & Zipser, D. (1989). A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2), 270-280.
17. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311-318).
18. Wang, H., Zhang, Y., & Yu, X. (2020). An overview of image caption generation methods. *Computational intelligence and neuroscience, 2020*(1), 3062706.
19. Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi, "Understanding of a convolutional neural network", IEEE – 2017