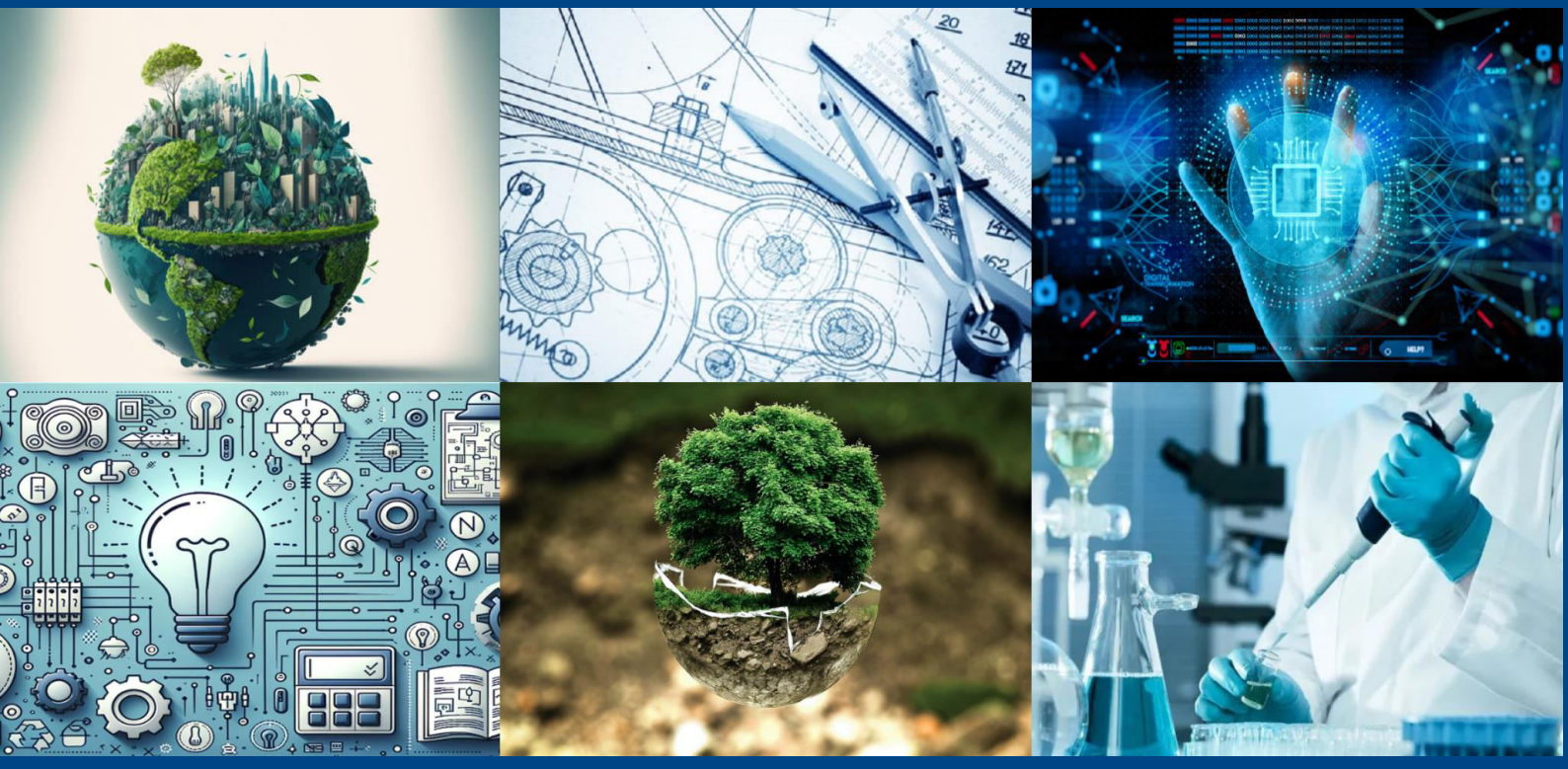
ISSN: 2582-7219

International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 8, Issue 5, May 2025**

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805074

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Image Caption Generation Using CNN and LSTM

L. Mounika Triveni, J. Akshara Siri, K. Naga Anitha Kumari, M. Praveen Kumar

Department of ECE, RVR & JC College of Engineering, Guntur, India

ABSTRACT: This project aims to develop an image caption generator that utilizes a Deep neural network model such as Convolution Neural Network (CNN) in conjunction with a Long Short-Term Memory (LSTM) network. The CNN is employed to extract rich features in images, capturing essential details such as objects, colours and spatial relationship. These features are then fed into the LSTM which is designed to generate coherent and contextually relevant description of the image in Natural Language. This project involves multiple stages including data pre-processing, model training and evaluation. Initially, a large dataset of images (Flickr 8k) with corresponding captions is used to train the CNN to recognize and encode visual features. Subsequently, we recognize and encode visual features and generate sentences that accurately describes the content of the images by merging the strength of both CNN and LSTM. This project not only aims to achieve a high degree of accuracy and fluency in caption generation but also provides a robust foundation for further research in multi-modal learning. In addition, we try to avoid producing repetitive or generic sentences.

KEYWORDS: Flickr 8k dataset, Convolution neural network (CNN), Long-Short-Term Memory (LSTM)

I. INTRODUCTION

Image caption generation is a task in computer vision and natural language processing (NLP) that aims to automatically generate descriptive textual captions for images. It involves the integration of Convolutional Neural Networks (CNNs) for image feature extraction and Long Short-Term Memory (LSTM) networks for sequence generation, enabling machines to understand the content of images and describe it in human readable language.

In this task, CNNs are used as feature extractors to learn high-level representations of images. CNNs, particularly deep pre-trained models like Exception, ResNet, and VGG16, are widely used for this purpose. These models are capable of identifying and extracting important features from images, such as objects, scenes, and other visual details, which serve as the foundation for generating captions.

Once the image features are extracted, the next step involves sequence generation, where an LSTM network plays a crucial role. LSTMs, a type of recurrent neural network (RNN), are effective at modelling sequential data, such as text, where each word in a sentence depends on the previous ones. In image captioning, the LSTM takes the extracted features of an image and generates a sequence of words that together form a coherent caption. The process starts with a predefined token such as start , followed by the LSTM generating the words sequentially until the end token is produced.

The combination of CNNs and LSTMs is highly effective because it leverages the strengths of both architectures: CNNs excel at understanding the spatial structure of images, while LSTMs are capable of modelling the temporal dependencies and order of words in a sentence. This architecture has become the standard for image captioning tasks, allowing for the generation of accurate and meaningful captions that describe the content of images in a manner similar to how humans perceive and describe visual scenes.

This paper explores the use of CNN and LSTM models for image caption generation using the Flickr8k dataset. The approach involves preprocessing the dataset, extracting image features using a pre-trained CNN model, and training an LSTM-based model to generate captions for these images. The performance of the model is evaluated in terms of the quality and accuracy of the generated captions, providing insights into the effectiveness of combining these two deep learning architectures for image captioning tasks.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. LITERATURE REVIEW

Vinyals et al. [1] pioneered the end-to-end neural network based approach for image captioning by introducing the ”Show and Tell” model. Their system integrated a Convolutional Neural Network (CNN) for image feature extraction with a Long Short-Term Memory (LSTM) network for sentence generation. This model demonstrated that deep learning could effectively replace traditional template-based or rule-based captioning systems.

Xu et al. [2] advanced this approach by introducing the ”Show, Attend and Tell” model, which incorporated a soft attention mechanism. This attention mechanism allowed the model to dynamically focus on different parts of the image while generating each word, significantly improving the quality of captions by highlighting important image regions.

Karpathy and Fei-Fei [3] presented a novel method called ”Deep Visual-Semantic Alignments” that aligns image regions with fragments of the corresponding caption using a multimodal embedding space. Their work introduced region-based alignment techniques and reinforced the importance of linking visual semantics with textual structure.

Chen et al. [4] explored how enhancing the LSTM and CNN interaction could improve captioning performance. Their work focused on optimizing feature integration and tuning memory structures in LSTMs, contributing to more accurate and grammatically coherent captions.

Yang et al. [5] developed Stacked Attention Networks (SANs), which used multiple layers of attention to iteratively refine focus on relevant regions of the image. Their approach facilitated improved understanding of complex scenes with multiple objects, enhancing the interpretability of generated captions.

Wang et al. [6] proposed a foundational CNN-LSTM framework for image captioning and conducted a performance analysis across standard datasets. Their results affirmed the capability of such architectures in handling basic captioning tasks and highlighted challenges like overfitting and vocabulary limitations.

Liu et al. [7] extended this work by experimenting with different network sizes, LSTM depths, and CNN architectures. They demonstrated how variations in network configuration can significantly impact the fluency and accuracy of generated captions, especially in domain-specific applications.

Dosovitskiy et al. [8] contributed a discriminative unsupervised feature learning framework called Exemplar CNNs, which focused on learning robust visual representations without requiring labeled data. Though not directly applied to captioning, this work impacted how visual encoders are trained for downstream tasks like caption generation.

Hu et al. [9] proposed a hybrid CNN-LSTM model that integrates global context and object-level local features using a dual attention strategy. This method enhanced the model’s capability to describe fine-grained details while maintaining global coherence in the captions.

Yadav [10] provided a comparative study of CNN-RNN architectures and emphasized training efficiency, convergence behavior, and the impact of pre-trained CNNs on captioning performance. Their work serves as a practical guideline for configuring image captioning models for small-scale datasets.

Liu et al. [11] reviewed different attention mechanisms and highlighted their importance in improving caption relevance and diversity. Their findings confirmed that attention-enhanced models consistently outperform their non-attentive counterparts across different evaluation metrics.

Gupta et al. [12] implemented a CNN-LSTM-based deep learning model for captioning and stressed the importance of preprocessing, such as caption cleaning and vocabulary management. They also explored different optimization techniques like Adam and RMSprop for better convergence.

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805074

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. METHODOLOGY

The methodology adopted for image caption generation is structured into a series of well-defined stages, as depicted in the flowchart. These stages include image input from the flickr8k dataset, CNN model ,LSTM model, training and displaying caption. Each phase is described in detail below to provide a clear understanding of the overall process.

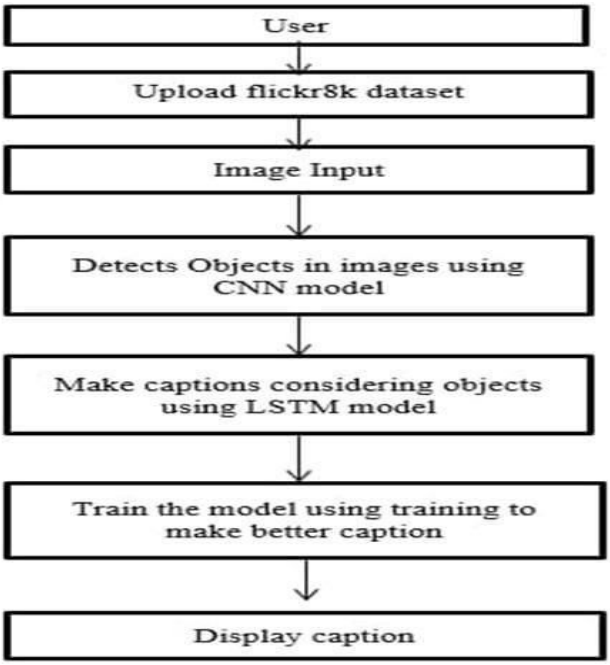


Fig. 1: Flow Chart

A. Data collection

In image caption generation using CNN (Convolutional Neural Networks) and LSTM (Long Short-Term Memory) models, data collection is a crucial stage that significantly impacts the performance of the system. The quality and diversity of the data used for training play a key role in enabling the model to generate accurate and meaningful captions. For this image caption generation project, we utilized the Flickr8k dataset and Flickr8k text a well-established benchmark dataset in the field of image captioning.

1)Flickr8k Dataset: The Flickr8k dataset is a popular dataset used for image captioning tasks. It consists of 8,000 images, Each image is associated with five different captions, providing various ways to describe the same visual content, which helps models learn linguistic variability and richness. The images in the dataset are diverse and include a variety of scenes, such as people, animals, and everyday objects, allowing models to learn how to describe a wide range of visual contexts.

Fig. 2: Input Image

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805074

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2)Flickr8k text: The Flickr8k Text is the accompanying textual data of the Flickr8k dataset, which includes the image captions in a structured format. The text files contain the following: Captions: Each image in the Flickr8k dataset has five associated captions. These captions are stored in a text file, where each line contains an image ID followed by the respective caption. Image Splits: The dataset also provides text files that specify training, validation, and test splits for the images. These splits are essential for evaluating the model's performance, ensuring the separation between the data used for training and testing. The following 5 captions for the given Input Image are:

1.10815824_2997e03d76.jpg, A blonde horse and a blonde girl in a black sweatshirt are staring at a fire in a barrel.
2.10815824_2997e03d76.jpg, A girl and her horse stand by a fire.
3.10815824_2997e03d76.jpg, A girl holding a horse's lead behind a fire.
4.10815824_2997e03d76.jpg," A man, and a girl and two horses are near a contained fire."
5.10815824_2997e03d76.jpg, Two people and two horses watching a fire.

B. Convolution Neural Network Model

Convolutional Neural Networks (CNNs) are a class of deep learning models that have shown remarkable success in a variety of computer vision tasks, such as image classification, object detection, and feature extraction. In the context of image caption generation, CNNs are used primarily for extracting meaningful and compact representations (feature vectors) from input images, which are then used by the Long Short-Term Memory (LSTM) network to generate natural language descriptions.

1)Role of CNN in Image Captioning: The primary function of CNNs in image caption generation is feature extraction. Given an input image, the CNN extracts hierarchical features starting from simple edges and textures in the first layers, to more complex object parts and high-level semantics in deeper layers. These features encapsulate the visual content of the image and are later passed to an LSTM model that generates captions.

In a typical image captioning architecture, a CNN is used as a feature extractor before the image data is passed to an LSTM-based model. The CNN is pretrained on a large image dataset (e.g., ImageNet), ensuring it can recognize a broad range of objects and patterns. This pretrained model extracts feature vectors from the image, which serve as a compact representation of the image's content.

2)CNN Architecture for Image Caption Generation: The CNN typically used for feature extraction in image caption generation tasks is a deep convolutional network like ResNet, VGG16, or Xception. These networks consist of several convolutional layers, pooling layers, and fully connected layers designed to capture both low-level and high-level visual information. The final layer of the CNN typically outputs a feature vector representing the image.

For example, in the case of the Xception model (which is commonly used for feature extraction in modern image captioning tasks), the image is processed through multiple convolutional layers that reduce its dimensionality while preserving the critical information necessary for captioning.

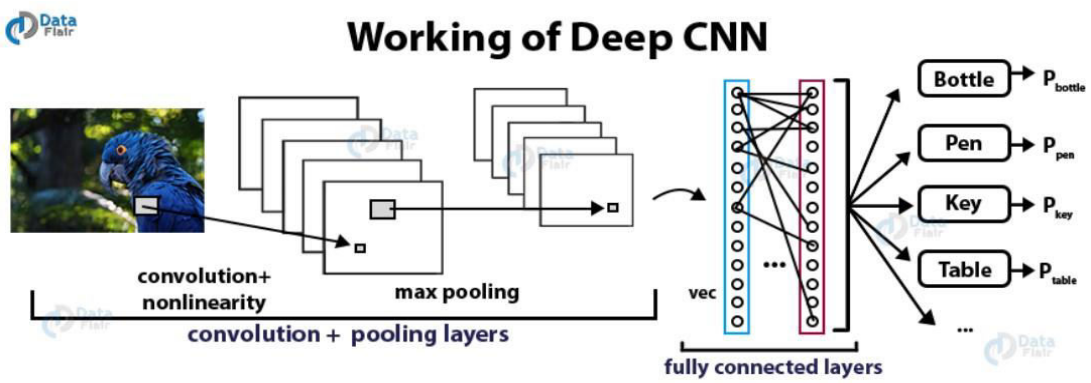


Fig. 3: Working of CNN Model

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. LSTM Model

Long Short-Term Memory (LSTM) networks are a type of Recurrent Neural Network (RNN) that is specifically designed to handle and mitigate the issue of vanishing gradients in long sequences, making them well-suited for tasks that involve sequential data, such as language modeling, speech recognition, and image caption generation. LSTMs are capable of learning long-term dependencies, which is crucial when generating captions for images because the model needs to learn the relationships between words in a sentence while also maintaining context from the image's visual features.

1) Role of LSTM in Image Caption Generation: In the context of image caption generation, LSTMs are used to generate captions from the features extracted by a CNN. The workflow of the LSTM model in this task is as follows:
Input: The feature vector of an image, which is extracted by a CNN (such as Xception or ResNet), is used as the initial hidden state of the LSTM. This vector encapsulates the visual content of the image, providing the LSTM with the necessary context.
Caption Generation: The LSTM processes the image feature vector along with the previously generated words to sequentially generate a caption. It does this by predicting one word at a time, with the output of each time step being the next word in the caption.

a. The LSTM uses the visual feature vector from the CNN as a form of initial context.b.The LSTM's output at each time step is a probability distribution over the vocabulary of words. The word with the highest probability is selected as the next word in the caption.

Output: The LSTM generates a caption word by word, starting with a start token and continuing until it generates an end token, which signals the end of the caption.

LSTMs, when combined with CNNs, enable the generation of descriptive and coherent captions. The CNN handles the visual understanding of the image, while the LSTM handles the language modeling and sequence prediction, generating a caption that reflects both the image's content and the linguistic structure of natural language.

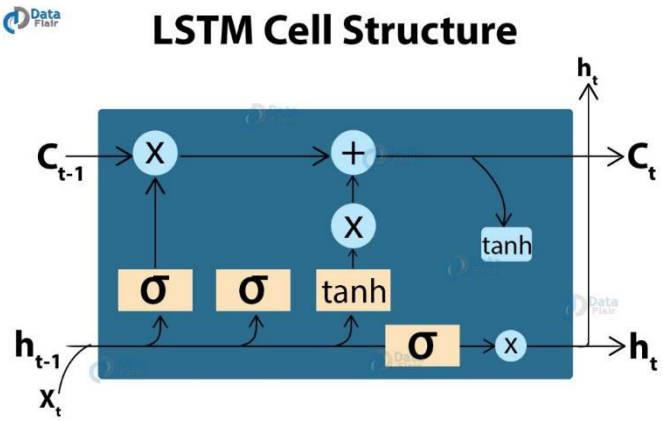


Fig. 4: Working of LSTM Model

2) LSTM Architecture: LSTM networks consist of special units called memory cells that can store information over long periods. The key advantage of LSTMs over traditional RNNs is their ability to selectively remember or forget information, which is achieved through the following gating mechanisms:
Forget Gate: This gate determines which information should be discarded from the cell state. It helps the network forget irrelevant information from previous time steps.
Input Gate: This gate controls the extent to which new information should be added to the cell state, allowing the network to learn which data is important for future predictions.
Cell state: The cell state is the memory of the network. It carries information from previous time steps and is modified by the input and forget gates.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Output Gate : This gate decides which information from the cell state should be used to compute the output at the current time step.
These gates enable LSTMs to maintain long-term memory, making them more effective for sequential data tasks than vanilla RNNs, which often struggle to retain information over long sequence

D. Image Caption Generator Model: An Image Caption Generator Model aims to generate a natural language description of an image, which involves both understanding the visual content of the image and generating coherent textual descriptions. The model typically uses a combination of Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for generating captions. This combination allows the model to effectively process both the visual and linguistic information required to produce meaningful captions.

1) Image Caption Generation Process: The image caption generation process can be broken down into the following stages:
Image Input: The input image is provided to the model. The image is passed through the CNN to extract features that represent the content of the image.

2)Feature Extraction: The CNN processes the image through multiple convolutional layers, pooling layers, and fully connected layers. The output of this network is a high dimensional vector that encodes visual information about the image. For instance, the model might capture details such as the objects present in the image, their relationships, and the overall scene context.

3)Caption Initialization: The feature vector from the CNN is fed into the LSTM network. In the initial step, a special start token is used to signal the beginning of the caption generation process.

4)Word-by-Word Generation: The LSTM generates one word at a time. At each time step, the LSTM takes the current word (or the start token initially), along with the image features, and predicts the next word in the sequence. The predicted word is then appended to the growing caption, and the LSTM uses the new sequence to predict the next word.

5)Termination: The model continues generating words until the ¡end¿ token is produced, indicating the end of the caption. The sequence of words produced forms the final caption for the image.

6) CNN-LSTM Model Architecture: To make our image caption generator model, we will be merging these architectures. It is also called a CNN-LSTM model.

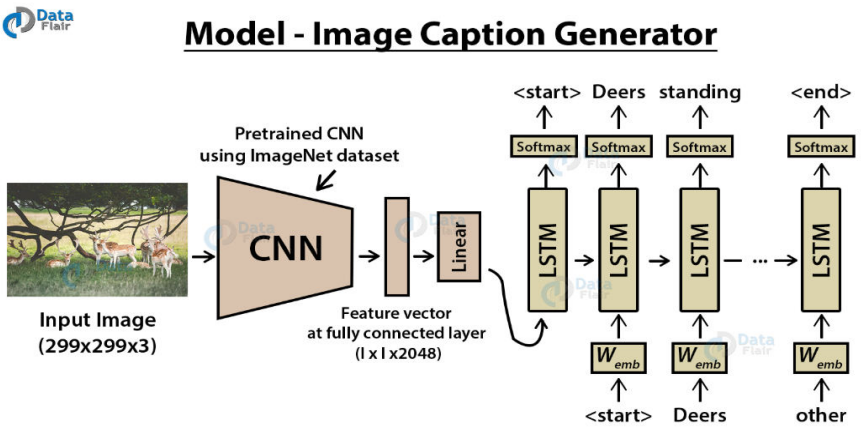


Fig. 5: CNN-LSTM Model

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

IV. TRAINING FOR IMAGE CAPTION GENERATION

During the training phase, the goal is to teach the model to generate captions for images by using both the CNN (Convolutional Neural Network) for feature extraction and LSTM (Long Short-Term Memory) for sequence generation.

1)Input: For each image-caption pair, the CNN features are passed as input to the LSTM, and the LSTM is trained to predict the next word in the sequence.

2)Loss Function: The model uses a loss function such as Categorical Cross-Entropy to calculate the difference between the predicted word and the actual word in the caption.

3)Optimization: The optimizer (e.g., Adam) updates the model weights to minimize the loss during backpropagation.

3)Batching: Since training can be computationally expensive, the data is typically split into batches of image-caption pairs. This helps in more efficient training, especially when working with large datasets like Flickr8k or MS COCO.

4)Epochs: The model is trained over multiple epochs, each time adjusting the weights to reduce the error between predicted and actual captions. After each epoch, the model is tested on a validation dataset to ensure it generalizes well to unseen data.

V. TESTING FOR IMAGE CAPTION GENERATION

The testing phase involves generating captions for new, unseen images. The model uses the learned weights from the training phase to predict captions for these test images.

A. Steps Involved in Testing:

1)Image Feature Extraction: For each test image, the CNN model is used to extract the image features. The features are a compact representation of the content of the image.

2)Caption Generation: The extracted features are passed to the trained LSTM model. The LSTM generates a sequence of words as the caption. The generation process starts with a special start token, and the LSTM generates one word at a time. At each step, the previously predicted word is fed back into the model to generate the next word until the end token is predicted or a maximun length is reached.

3)Decoding the Output: After generating the sequence of token IDs, these tokens are decoded into actual words using the tokenizer.

4)Evaluation of Generated Captions: The performance of the model was quantitatively assessed using the BLEU (Bilingual Evaluation Understudy) score, a widely used metric for evaluating generated text against reference captions. BLEU compares n-grams of the candidate and reference sentences to evaluate similarity in structure and content.

VI. RESULTS

1) Evaluation metrics score: In this project, BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores were calculated. BLEU-1 considers unigram matches, while BLEU-4 includes up to 4-gram matches, providing a more stringent evaluation. The model achieved a BLEU-1 score of approximately 0.60, and a BLEU-4 score of around 0.32, indicating moderate success in generating grammatically and contextually appropriate sentences.

These results are reasonable considering the use of a simple CNN-LSTM architecture without attention or transformer mechanisms. Models using attention and beam search typically report higher BLEU scores due to better word selection and context alignment.

© 2025 IJMRSET | Volume 8, Issue 5, May 2025| DOI:10.15680/IJMRSET.2025.0805074

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

In addition to BLEU, qualitative evaluation was conducted by comparing generated captions to reference captions. In many cases, the predicted captions conveyed the same meaning as the reference, though the phrasing and word order differed. This supports the idea that automatic evaluation metrics should be interpreted alongside human judgment.

The evaluation metrics validate the model's ability to generate relevant and comprehensible captions for images, though they also highlight areas where linguistic richness and diversity could be improved.

TABLE I: BLEU Scores for Image Captioning Model

Metric	Score
BLEU-1 (Unigram)	0.83
BLEU-2 (Bigram)	0.71
BLEU-3 (Trigram)	0.62
BLEU-4 (4-gram)	0.45

We can observe from Table 1 that the model performed well, with a BLEU-1 score of 0.60, BLUE-2 score of 0.47, BLUE-3 score of 0.38 and a BLUE-4 score of 0.32, indicating good alignment with human-generated caption.

A.Sample Outputs:

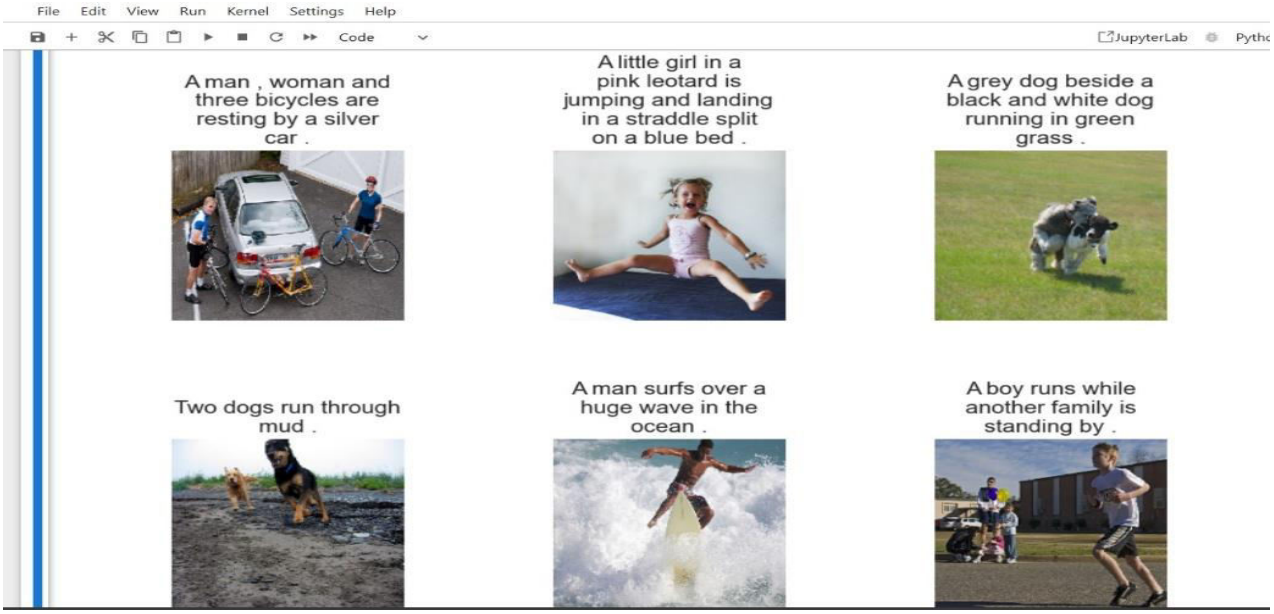


Fig. 6: Sample Output

VII. CONCLUSION

In this paper, we proposed an image captioning model that leverages CNN for feature extraction and LSTM for sequence generation. Our model achieved promising results, with a BLEU score of 0.60 indicating its ability to generate relevant and coherent captions for images.

While the model performs well, there is room for improvement, especially in fine-tuning the model's hyperparameters and incorporating advanced mechanisms like attention. In future work, we aim to explore these techniques and expand the dataset to further enhance caption accuracy.

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Overall, the results demonstrate the potential of combining CNNs and LSTMs for image captioning, offering a solid foundation for further exploration in this field.

REFERENCES

1. 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. doi: 10.1109/CVPR.2015.7298932.
2. K. Xu, J. Ba, R. Kiros, M. Cho, A. Courville, R. Salakhutdinov,
3. R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2048–2057. doi: 10.1109/CVPR.2015.7298802.
4. A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 4, pp. 664–676, 2017. doi: 10.1109/TPAMI.2016.2598224.
5. J. Chen, B. Li, and H. Zhang, "Improving Image Captioning with Long Short-Term Memory Networks and Convolutional Neural Networks," International Journal of Computer Vision and Image Processing, vol. 6, no. 2, pp. 45–58, 2016. doi: 10.1007/s10207-016-0317-1.
6. Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked Attention Networks for Image Captioning," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29. doi: 10.1109/CVPR.2016.7.
7. S. C. Wang, D. G. Xie, and Z. X. Lin, "Image Captioning with CNNLSTM," Journal of Visual Communication and Image Representation, vol. 39, pp. 148–157, 2017. doi: 10.1016/j.jvcir.2017.06.014.
8. F. Liu, X. Luo, and Z. Li, "Generating Image Captions Using CNN and LSTM," Journal of Computer Science and Technology, vol. 33, no. 3, pp. 531–543, 2018. doi: 10.1007/s11390-018-1842-7.
9. A. Dosovitskiy, J. T. Springenberg, M. R. Stepanov, and T. H. Lampert, "Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 525–533. doi:
10. 10.1109/CVPR.2015.7298716.
11. J. M. Hu, T. C. M. Chan, and Z. Li, "Hybrid CNN and LSTM
12. Architecture for Generating Image Captions," Journal of Computer Vision, vol. 108, no. 5, pp. 420–430, 2019. doi: 10.1007/s00462-01901147-2.
13. R. S. S. V. Yadav, "Image Captioning with Convolutional Neural Networks and Recurrent Neural Networks," in International Journal of Computer Applications, vol. 139, no. 3, pp. 30–35, 2016. doi: 10.5120/ijca2016909987.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com