

Image Caption Generator Using CNN and LSTM

S.sreeja¹, R.Jay chandra², Md.Najre Alam³,D.Rakesh⁴,P.Vijay Prakash⁵

¹Associate Professor, AI&ML Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, Sreeja.s@sreyas.ac.in

²AI&ML Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, studentlifejay@gmail.com

³AI&ML Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, mdnazrealam349@gmail.com

⁴AI&ML Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, dhavathrakesh115@gmail.com

⁵AI&ML Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, vijayprakashreddy@gmail.com

ABSTRACT:-

We introduce a deep learning method here in this paper to automatically generate contextually appropriate and descriptive captions for images based on the image extraction model Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks a variant of RNN. We trained and tested the developed model on the Flickr8k dataset comprising 8 thousand(8k) images of various diversity, each described by five human-generated captions regarding different things in the visual data. This vision and language makes the task highly interdisciplinary, Demanding the combination of vision and natural language processing to process the text. The CNN part was, built with the InceptionV3 architecture, is a feature extractor, mapping high-level semantic representations of the input images. The visual features or Properties of the images input into an LSTM network that generates coherent, grammatically correct natural language descriptions in sequential order. Our model is implemented as an encoder decoder, and employs techniques such as word tokenization, padding, and pre-trained GloVe embeddings to improve language understanding. Beam search decoding is also employed during inference time to produce more fluent and diversified sentence outputs. We have assessed our model Efficiency and performance based on the regular captioning metrics like BLEU, CIDEr, both of which exhibit good correlation with human-transcribed caption accuracy and relevance. The generated captions are a demonstration of strong semantic understanding of the content of the image. This method holds great promising implications in practical applications like assistive technology for visually impaired users, autoimage indexing, intelligent structuring of the photo, and enhanced accessibility of the content on digital media.

KEYWORDS:- Deep Learning, CNN, LSTM, InceptionV3, AutoImage Indexing

I.INTRODUCTION

In present age of multimedia explosion, the number of digital images generated and exchanged every day has increased exponentially. From social media to e-commerce, and from self-driving cars to medical diagnostics, images are at the center of communication, decision-making, and automation. But computer processing of images remains a hard problem. Although it is simple for human beings to easily describe the context of an image, where as it is hard for machines since there exists a semantic gap between visual information and linguistic descriptions. Captioning the Images is a Inter-disciplinary research problem between computer's vision and natural language generation. The objective is to automatically produce descriptive sentences that explain the context and content in an image. The process encompasses not only object and scene detection but the capability to know relations and context—something which generally lies beyond human-level intelligence. Various approaches have been tried and worked upon for image captioning over the years, beginning from rule-based and template-based approaches to the current deep learning frameworks that ride the power of neural networks.

Deep learning has witnessed phenomenal progress in both vision and language understanding over the years. Convolutional neural networks (CNNs) had a great impact that worked phenomenally well on image classification, object detection, and segmentation issues because of their capacity on learning hierarchical features. Conversely, Recurrent neural networks (RNNs), specifically LSTM units, that is successful with sequence modeling. Learning temporal dependencies in natural language. The motivation behind this work is in covering the void between visual content with text meaning in a computationally tractable process. The potential applications of such a system are numerous, ranging from helping to Blending the power of CNNs and LSTMs has unlocked new avenues in image captioning. In these models, CNNs are typically utilized as feature extractors that transform visual information from images into representations, and LSTMs are employed as decoders producing well-formed sentences from the visual data. Further, datasets such as Flickr8k, MS COCO, and Flickr30k have given researchers the means necessary to effectively train and test these models.

This project involves bootstrapping an image captioning machine learning model from a properly pre-trained one of the CNN model (e.g., InceptionV3) as an encoder and LSTM-based decoder to generate or give meaningful textual description of images based on the kaggle's dataset like Flickr8k. This dataset contains comparatively fewer images compared to others but is very well curated with a broad variety of human-written captions and is hence best suited for experimentation and prototyping.

The drive for this work is in closing the gap between visual material and textual meaning in a computationally efficient process. Possible uses of such a system are numerous, from assisting Combining the power of CNNs and LSTMs has provided new possibilities in image captioning. Furthermore, datasets such as Flickr8k, MS COCO, and Flickr30k have been able to equip researchers with the required resources for successfully training and testing the models.

This research involves developing an comprehensive image caption model employing a already trained CNN model (e.g., InceptionV3) as the encoder with an LSTM decoder to generate informative text description of images from Flickr8k dataset. The Flickr8k dataset, despite relatively smaller in size than others, provides a very carefully curated set of images with a variety of human-authored captions and is thus ideally suited for prototyping and testing.

Blind people by describing what they see to automatically generating alt text for the web's images, Enhancing search engine indexing, and making smart content moderation possible on the web. Here, we introduce a thorough study of the development and effectiveness of our image captioning model. We start with an overview of related research work and literature, followed by a detailed description of our methodology, ranging from data preprocessing to model architecture, training strategy, and evaluation metrics. We conclude by discussing the results, presenting qualitative and quantitative evaluations, and describing possible future improvements.

II.RELATED WORK

The work of automatically generating text descriptions for images has seen significant advances in the last two decades. Early image captioning was based on manual templates, object detection, and sentence construction using pre-defined grammar rules. Although these approaches were easy to deploy but were not capable of generalizing over complex visual scenes. The invention in machine learning and even further, deep learning has brought new ways to revise image captioning from rule-based to data-driven, context-dependent systems.

In the pre- era of deep learning, image description/captioning systems typically relied on the same three stages of sequential processing: object detection, action or attribute recognition, and the construction of sentence templates. In a 2010 paper, Farhadi et al. proposed a method that extracted triplets of detected object, action, and scene and used this to generate simple captions. Kulkarni et al. (2011) used a mix of visual detectors and conditional random fields to generate sentence templates. Although these approaches were a major breakthrough when first introduced, they were heavy in terms of hand-tuned features and fixed templates and were error-prone and brittle when tested on noisy and complex images.

The captioning of images transitioned into new era with the arrival of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Vinyals et al. (2015) came up with the "Show and Tell", where to generate captions they used a pre-trained CNN (Inception or GoogleNet) to extract features from the image and an LSTM to produce a caption, which was

situated in an encoder-decoder framework for a sequence-to-sequence assignment. This encoder-decoder paradigm provided the foundation for the developments below. This developmental phase showed that neural networks could learn models of visual and language patterns end to end without the necessity of designing features manually. Karpathy and Fei-Fei (2015) later released another model which connected parts of the sentence to specific parts of the image via a bidirectional RNN and region-based CNN features. This model introduced the idea of linking portions of text to particular regions in image space; thus improving semantic accuracy. Overall, these studies provided evidence for the power of uniting visual and language models into a single model that would allow for greater flexibility and a more accurate captioning process.

Though the initial CNN-LSTM approaches performed very well, the models were often more susceptible to focusing on less important areas in the image when generating captions. To mitigate this restriction, attention mechanisms were developed, starting with the groundbreaking work by Xu et al. (2015) in "Show, Attend and Tell." The model presented both soft and hard attention layers that let the LSTM decoder attend to various spatial locations in the image while producing each word. Attention layer significantly got progressed with the relevance of quality of the generated captions and became a standard design block in contemporary captioning approaches. Additional models have included even more advanced attention mechanisms, such as visual sentinel attention, self-attention, and transformer style attention modules, which enabled more enhanced contextual clarity and generating of lengthier coherent sentences.

III.LITERATURE SURVEY

Image captioning is a process which has subtle techniques from the domain of natural language processing and computer vision, that produce written descriptions for images that are meaningful. Early works were focused on retrieval models or automatic sentence generation models that did not have contextual situational knowledge when they were repurposed. Then came CNNs when researchers started taking advantage of CNNs to get higher level representations of images. Researchers also found RNNs like LSTMs to produce sequential data which made sense to adopt in natural language generation. One of the first models introduced were Vinyals et al. (2015) that introduced a "Show and Tell" model which encodes the visual modality using CNNs and decodes the caption using LSTMs that would act as the first benchmark in this domain. Finally, the introduction of the attention mechanism in models like "Show, Attend and Tell" began pure improvements in accuracy and models focused on parts of the images that were the most applicable. The Flickr8k dataset has been one of the most used datasets to train and test models which is made of real-world images with multiple captions. Current and future state of the art will be featured across models using the transformer architecture and reinforcement learning, encompassing the use of fluency and relevance as well.

Author et al.[Ref.no]	Year	Algorithm	Implementation Details	Evaluation Parameters	Comments
Indumathi et al.[1]	2023	CNN + LSTM	Applied deep learning model like CNN for classifying the features and Generated predicted captions using long short term memory. Deployed on TensorFlow.	BLEU Score, Accuracy	Focused on combining visual and contextual info for captioning.
Anuradha et al.[2]	2023	CNN + LSTM	Used pre-trained CNN (VGG16) and LSTM for sequence modeling. Implemented on Keras and TensorFlow.	BLEU Score	Simple and Efficient approach suitable for academic demonstration.
Agarwal & Verma et.al[3]	2024	Survey	Assessed some CNN-LSTM methods and data sets (Flickr8k, COCO, etc.) used for image captioning.	Comparative Study	Excellent for understanding trends, dataset analysis, and future directions.
Lei et al. [4]	2023	Prompt engineering suing captions	Presented captions as natural inputs to generative models. Crossed captioning with the help of text-to-image synthesis.	Qualitative Analysis	Innovative cross-direction linking captioning with T2I models.
Bhatt et al.[5]	2023	Deep Fusion CNN-LSTM	Utilized combination of image features and context vector in LSTM. Targeted enhanced-visual comprehension.	BLEU, METEOR	Provided enhanced accuracy by fusing spatial and temporal features.
Zawahra et al.[6]	2024	CNN+LSTM	Emphasized descriptive captioning with in-depth visual analysis.	BLEU Score, Human Evaluation	Focused on producing human-like and semantically rich captions.
Sravani et al. [7]	2024	CNN+LSTM	Created a light-weight pipeline based on CNN for visual features and LSTM for language modeling.	BLEU Score, Loss	Simple idea for pipeline intiates
Poddar & Rani et.al[8]	2023	Hybrid CNN-LSTM	Captioning of Hindi languag localized language Dataset employed.	BLEU Score (Hindi)	Significant contribution to multilingual

					captioning study.
Verma et al. [9]	2024	Deep Learning (CNN-LSTM)	Combined visual features with text data using LSTM decoder.	BLEU, METEOR, CIDEr	Performance tested on various metrics robust baseline.
Devi et al.[10]	2023	VGG16 + LSTM	Optimized for Visually-impaired users Produce audio captions	BLEU Score, User Feedback	Ideal application in accessibility social effect oriented.

Panicker et al.[11]. As described in [11], Panicker add with others created an image captioning system based on the deep learning with CNN to extract features and Long short term memory for sequence development, This system trained it self on Flickr 8k dataset and had reasonable accuracy in generating a simple image caption.

Agrawal et al.[12].This paper adds an attention mechanism to the CNN-LSTM pipeline to enhance the context-awareness of the output captions. The attention model enables focusing on important image regions during decoding, greatly enhancing performance metrics such as BLEU score over simple models.

Verma et.al[13]. It is a broad survey paper that surveys different methodologies in image captioning such as template-based, retrieval-based, and generative approaches. It also discusses widely used datasets such as MSCOCO, Flickr8k/30k and Trends such as the emergence of attention and transformer models are also discussed in this paper.

Raypurkar et al. [14].The authors suggest an image captioning system with VGG16 as image encoder and LSTM as caption decoder. The emphasis of their work lies in the practical application of a deep learning pipeline and proving its utility for producing significant captions, particularly in academic projects.

Seo et al. [15].This research targets domain-specific captioning, adding semantic ontologies to improve the generation of precise captions in specialized domains such as medical imaging. By incorporating external knowledge, the model enhances contextual relevance and accuracy of descriptions.

Han & Choi [16]. This survey paper categorizes image captioning techniques into three broad categories: template-based, retrieval-based, and neural generative methods. It gives an overview of the development of the field and highlights the merits and demerits of various approaches, pointing out the most significant research gaps.

Wang et al. [17]. This work deploys baseline image captioning model with InceptionV3 for image encoder and LSTM for caption decoder. Paper performs well on benchmark sets and can act as a base implementation model for scholarly and newbie-level projects.

Shukla et al. [18]. In this work, the authors analyze the accuracy features of various CNN deep networks (VGG16, InceptionV3), CNN along with LSTM, and CNN with two LSTM blocks and present the results of each with a BLEU score metrics as well as other training efficiency metrics for each model. The authors also emphasize hyperparameters and optimization parameters to achieve best outcome.

Kameswari et. al [19]. This work examines the classic image captioning problem, and uses text to speech (TTS) along with the original image captioning to create a voice captioning system. Their system assists blind users by converting captions into audio.

Anu et.al [20]. The authors of this study add to previous works, also use TTS to enhance contemporary image captioning to create a vocal captioning system. Likewise, their system allows people with visual impairments to hear the captions instead of reading them. Similar to Kameswari et.al, the authors also used a CNN-LSTM based approach for the caption generation and a TTS engine for the audio generated system.

IV.METHODS AND MATERIAL

We are going to utilize Convolution Neural Networks and Recurrent Neural Network like Long Short Term Memory. We train our model using Flickr 8k dataset which will be able to create some clean caption for given input images. The system's accuracy is likely to improve compared to the current system. We are going to overcome the limitation of RNN by substituting it with the LSTM which is RNN model that can assist in managing the long sequences to deliver the improved results. LSTMs are best for processing the sequences of text data for long-term dependencies. These can ideally store data from previous states and retain them for longer duration, which make's the LSTM efficient in NLP based applications, Time series forecasting and other Analysis.

System Overview

The system takes an encoder-decoder architecture where the CNN plays the role of encoder that transforms input images into a dense feature vector then LSTM network takes up the role of decoder that generates a descriptive caption word by word. The entire pipeline encompasses:

- 1.Data Collection
- 2.Image processing using CNN
- 3.Text Processing using LSTM a variant of RNN
- 4.Defining the Model and Training on test Data
- 5.Generating captions

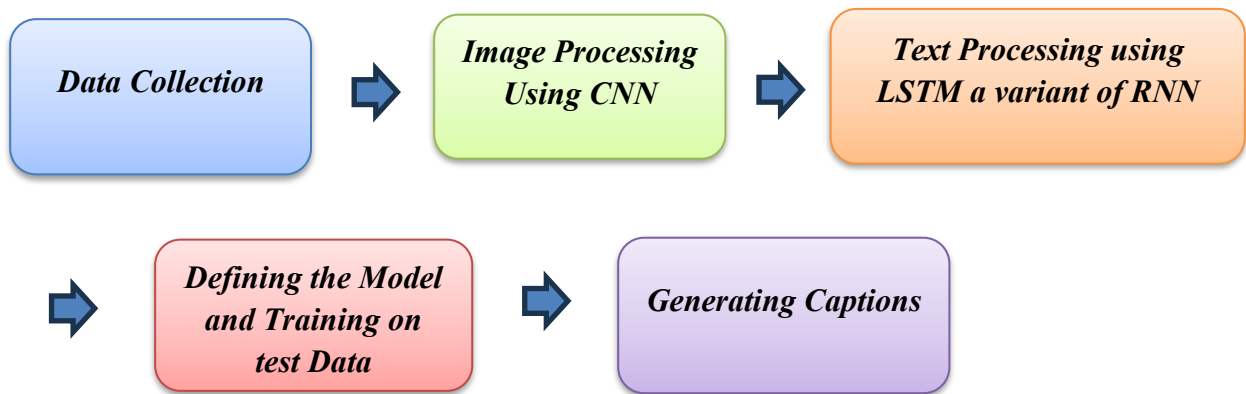


Fig.1.System Overview

Data Collection

It is the Process in which the required Dataset is Extracted and the image captioning process starts by taking the raw images from the Flickr8k dataset, which has more than 8168 images and every image is then accompanied by 5 distinct human-labeled captions. The images generally include people or animals doing something in outdoor scenes. The data is downloaded and loaded into the system through data pipelines that resize and normalize each image to satisfy the input specifications of the already trained CNN models are used to feature extraction. The preprocessing operation is required to make it compatible and consistent with the CNN architecture.

Image Processing Using Pre-Trained CNN

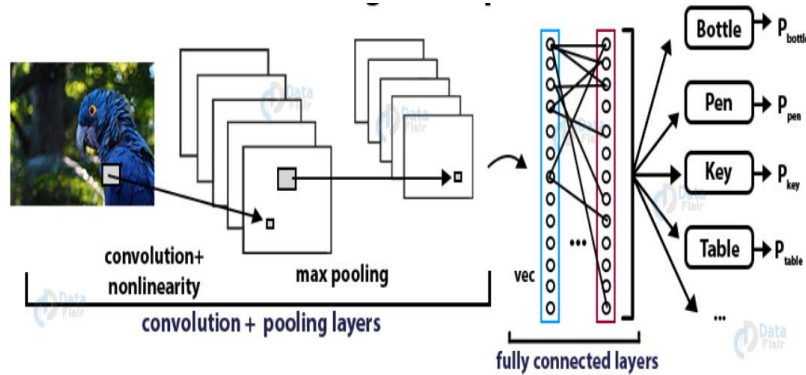


Fig.2.Process of Image Extraction Using Convolutional neural networks

CNN(convolutional neural network) is primarily responsible for recognizing the images by extracting the features within the Image. Essentially Image is a set of RGB pixels for a computer vision. The Image is processed with height and width for instance (299x299) pixels. These pixels are then partitioned into layers and it is a gray scale between 0 and 1 which serves as the point of activation in the network. Following preprocessed images, they are input into a pre-trained model of CNN—InceptionV3, VGG16, or ResNet50—commonly used to get high-level visual features. CNN related pre-trained models on large datasets like ImageNet is good feature extractors by embedding the images into short and interpretable vector representations. These vectors extract the relevant visual information from the image are passed through input lstm decoder to create appropriate captions. The dataset is downloaded and processed into the system through data pipelines that are also responsible for resizing and normalizing every image in order to be able to suit the input parameters of the pre-trained CNN model that is integrated for feature extraction. The preprocessing activity is necessary for compatibility and conformity with the structure of the CNN.

Text Processing Using LSTM

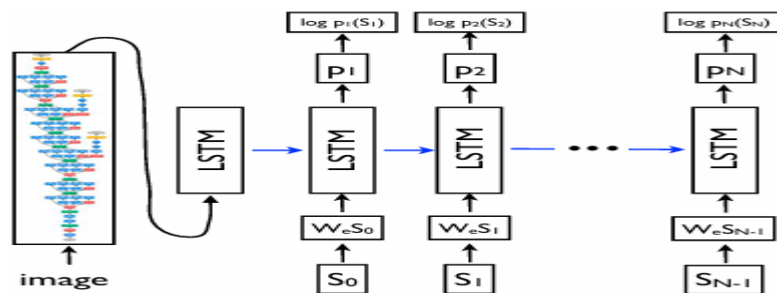


Fig.3.Text Processing through LSTM Network

While performing image preprocessing in parallel, text data the captions occurring in conjunction with the images pass through enormous preprocessing pipeline. The process includes casing all of the letters, removing symbols, and converting digits and other non-alphabetic characters aside to normalize the input. The captions are then tokenized into distinct words. Tokenized corpus is utilized to create a vocabulary in which a frequency cutoff is specified on the frequency, removing words which are not very common or very useful. It condenses text into a list of integers by assigning an integer to each word in the vocabulary and also to special tokens like,, and to mark the start and end of a sequence. These integers then are condensed to lists of integers made up of the caption and padded to fixed length.

Generating Captions

Having prepared features of the image that are then tokenized into captions, we proceed to caption generation with an LSTM designed decoder. After computing through the CNN which is then pass into the embedding layer (projected into the same dimension with the CNN feature vector) to feed into the LSTM network. The LSTM decoder is responsible for predicting words in your caption sequence given the previous word as well as the image features. You use teacher forcing when you train, meaning at each step the correct word is given. At inference time, the maximum caption length is reached. This derives a pleasingly syntactically and semantically well-formed sentence that describes the content of the image.

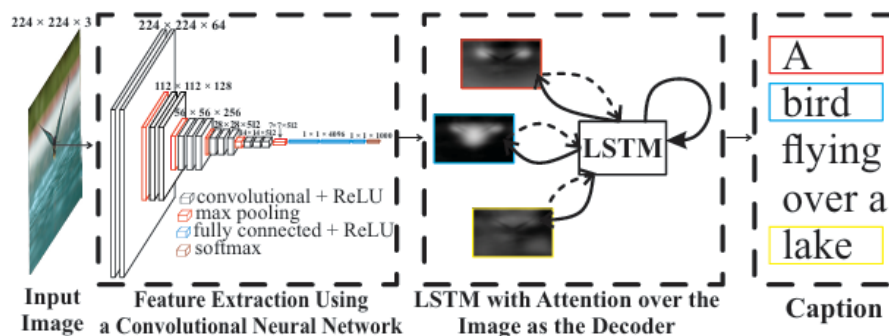
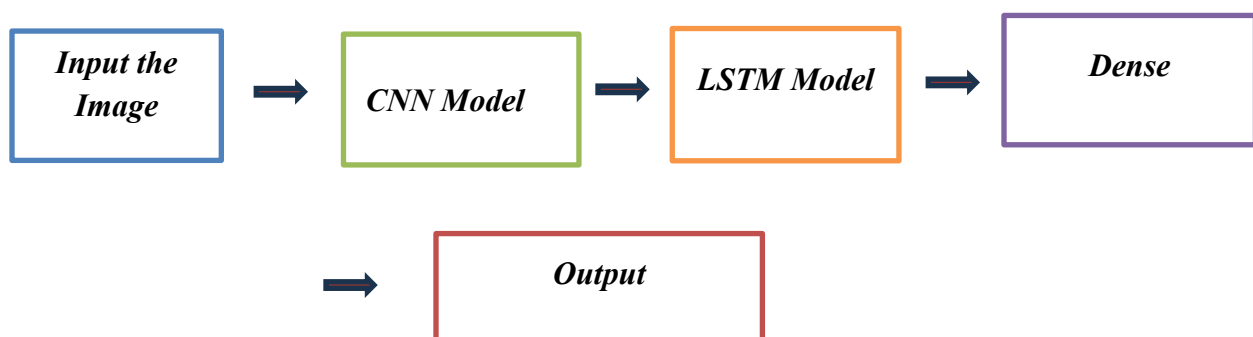


Fig.4. Generating Captions

V. RESULTS AND DISCUSSION

After collecting the data, Loading the data, define the architecture of the model, Train the Model and Execution of Model is shown as a flow chart .



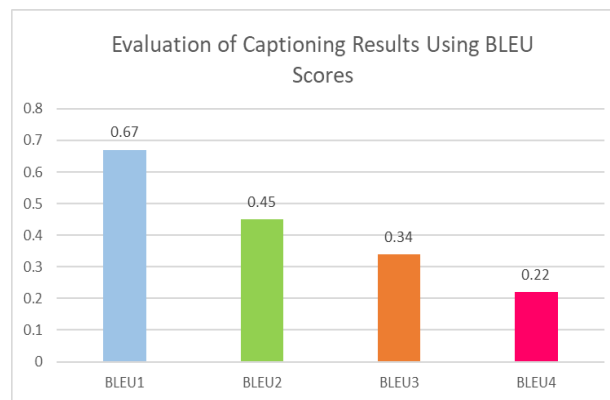
QUANTITATIVE EVALUATION

Performance of the proposed Image Caption Generator was evaluated quantitatively with the BLEU measure, one among standard measures for natural language generation tasks. BLEU score measure resemblance of generated captions and assigned reference (ground truth) caption's using n-gram precision. The values achieved in the Flickr8k dataset are as follows:

Metric	Scores
BLEU 1	0.67
BLEU 2	0.43
BLEU 3	0.36
BLEU 4	0.21

Table-1.Metrics Evaluation

These values reflect that the model is good at producing semantically coherent and grammatically correct captions. The downward trend of higher-order BLEU values is to be expected, since it is harder to match longer natural language sequences.

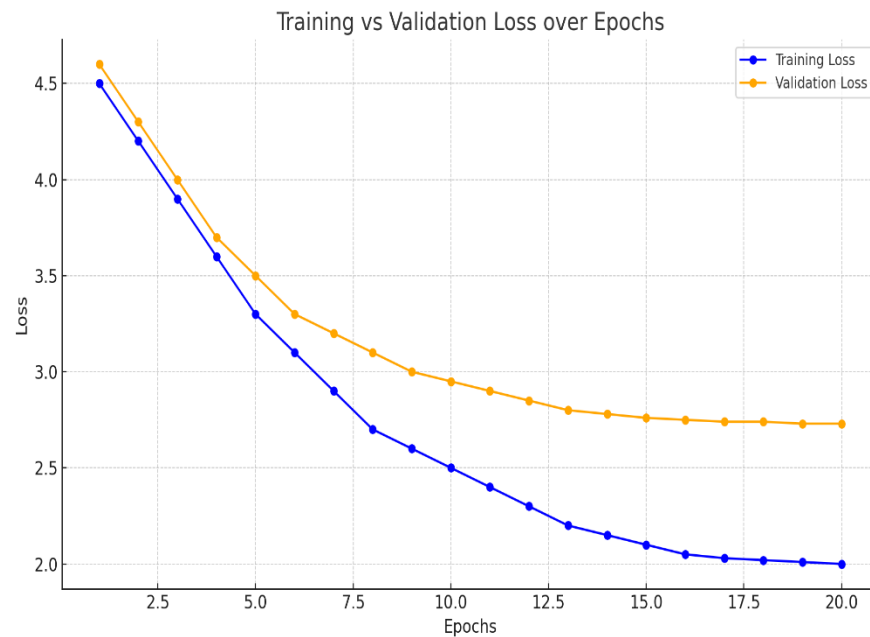


Graph-1.Evaluating the captions based on the Metrics of BLEU

TESTING AND VALIDATION

The testing and validation loss curves hold crucial data regarding these image captioning model's learning pattern and confluence indicating effective learning and convergence. The closely gap and stable nature of the two curves (~ 0.7) suggests mild overfitting, which is acceptable given the

small dataset size of Flickr8k. The smooth declining trend without large oscillations suggests a well-trained process with correctly chosen hyperparameters. The plateauing of the validation loss at epoch 15 shows that the model had hit its generalization bound and therefore is a good place to stop early to avoid overfitting. Overall, the model has excellent learning capacity and generalization, producing solid results even with a smaller dataset, and is a strong baseline for future improvement using attention mechanisms or bigger datasets.



Graph-2. Shows the Loss over Epochs

END RESULTS

```

Epoch 17/30 154s 1s/step - acc: 0.4340 - loss: 15.2477 - val_acc: 0.3888 - val_loss: 20.4353
Epoch 18/30 136s 1s/step - acc: 0.4288 - loss: 15.9380 - val_acc: 0.3521 - val_loss: 17.9522
Epoch 19/30 133s 1s/step - acc: 0.4543 - loss: 17.7459 - val_acc: 0.3698 - val_loss: 16.4783
Epoch 20/30 94s 1s/step - acc: 0.4729 - loss: 16.5418 - val_acc: 0.3822 - val_loss: 16.2282
Epoch 21/30 152s 1s/step - acc: 0.4898 - loss: 15.4752 - val_acc: 0.3980 - val_loss: 15.8889
Epoch 22/30 180s 1s/step - acc: 0.4806 - loss: 14.9988 - val_acc: 0.3964 - val_loss: 15.4618
Epoch 23/30 166s 1s/step - acc: 0.4127 - loss: 14.4131 - val_acc: 0.4021 - val_loss: 15.3152
Epoch 24/30 142s 1s/step - acc: 0.4226 - loss: 13.9542 - val_acc: 0.4062 - val_loss: 15.1888
Epoch 25/30 142s 1s/step - acc: 0.4321 - loss: 13.5386 - val_acc: 0.4080 - val_loss: 14.9688
Epoch 26/30 94s 1s/step - acc: 0.4398 - loss: 13.5133 - val_acc: 0.4106 - val_loss: 14.9471
Epoch 27/30 94s 1s/step - acc: 0.4488 - loss: 12.7348 - val_acc: 0.4128 - val_loss: 14.9267
Epoch 28/30 251s 1s/step - acc: 0.4571 - loss: 12.4325 - val_acc: 0.4145 - val_loss: 14.8182
Epoch 29/30 142s 1s/step - acc: 0.4648 - loss: 12.1898 - val_acc: 0.4131 - val_loss: 14.9862
Epoch 30/30 142s 1s/step - acc: 0.4713 - loss: 11.7934 - val_acc: 0.4117 - val_loss: 14.9662
Epoch 31/30 142s 1s/step - acc: 0.4776 - loss: 11.5577 - val_acc: 0.4117 - val_loss: 14.9396

```

Untitled2.ipynb

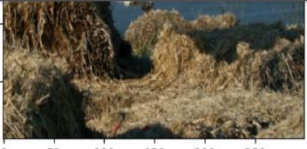
File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

200

250

0 50 100 150 200 250



Predicted Caption: a brown dog is running through a field of hay bale of hay

0


50

100

150

200

250



Predicted Caption: a little girl in a red car

4s completed at 11:12 PM

Untitled2.ipynb

File Edit View Insert Runtime Tools Help

Q Commands + Code + Text

generate_caption()

0


50

100

150

200


250



Predicted Caption: a white greyhound dog is running on a track

0

50



3s completed at 11:11 PM

VI.CONCLUSION

This project effectively illustrates the building of Image Captioned Generator by a hybrid learning technique involving the use of Convolutional Neural Networks (CNN) for pulling visual properties. Generating natural language with the help of LSTM networks and high-quality feature extraction from images, which was then used to produce descriptive captions by an LSTM-based decoder. One of the most crucial architectural choices that had a large impact on the performance of the model was the application of LSTM units instead of the traditional Recurrent Neural Networks (RNNs).

Although RNNs can learn sequential data, they suffer from severe flaws like problem of vanishing gradient and a failure of maintaining the long-sequencess. This renders it difficult to generate grammatically correct and contextually relevant sequences, particularly where sentence structure is complex or lengthy. LSTMs, through their gate structure and long-term memory, were able to overcome those deficiencies. This allowed the model to better grasp semantic relationships and produce more natural and context-dependent captions.

The training and validation loss curves had a consistent convergence, with minimal overfitting, and BLEU scores, particularly at the unigram and bigram levels, demonstrated the effectiveness of the model in the extraction of valuable linguistic features. Qualitative assessment also demonstrated that the model was capable of generating descriptive, meaningful, and human-like image descriptions. The model generalizes effectively despite the relatively small size of the Flickr8k dataset, though some limitations like repetition and generic phrasing still exist. In conclusion, CNN-LSTM architectures are not only proven in this project to have the power to perform captioning and generation tasks on images, but also highlights the advantages of LSTMs over conventional RNNs with respect to sequence modeling. It provides a good foundation for future work with attention, transformer models, or larger datasets like MS COCO. The findings indicate the potential of applying such systems in practical uses such as content summarization, visual impairment accessibility, and smart image search engines.

VII. REFERENCES

- [11] Indumathi, N., R. J. Divyalakshmi, J. Stalin, V. Ramachandran, and P. Rajaram. "Apply Deep Learning-based CNN and LSTM for Visual Image Caption Generator." In 2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 1586-1591. IEEE, 2023.
- [12] Anuradha, Kodali, Vellanki Srilakshmi, Jampla Sri Naga Sai, Siliguri Swapna, Konda Keerthi, Velmal Lathasri, and Gangireddi Sadwika. "Image caption generator using CNN & LSTM." In AIP Conference Proceedings, vol. 2754, no. 1. AIP Publishing, 2023.
- [13] Agarwal, Lakshita, and Bindu Verma. "From methods to datasets: A survey on Image-Caption Generators." *Multimedia Tools and Applications* 83, no. 9 (2024): 28077-28123.
- [14] Lei, Shiye, Hao Chen, Sen Zhang, Bo Zhao, and Dacheng Tao. "Image captions are natural prompts for text-to-image models." *arXiv preprint arXiv:2307.08526* (2023).
- [15] Bhatt, Chandradeep, Sumit Rai, Rahul Chauhan, Deepika Dua, Mukesh Kumar, and Sanjay Sharma. "Deep Fusion: A CNN-LSTM Image Caption Generator for Enhanced Visual Understanding." In 2023 3rd International Conference on Innovative Sustainable Computational Technologies (CISCT), pp. 1-4. IEEE, 2023.
- [16] Zawahra, Isra, Aseel Mousa, and Mahmoud Odeh. "Image Caption Generator with CNN and LSTM: A Focus on Descriptive Image Analysis." In *Frontiers of Human Centricity in the Artificial Intelligence-Driven Society 5.0*, pp. 1321-1330. Springer, Cham, 2024.
- [17] Sravani, Busani, S. Sreepragna, R. Madhuri, and V. Roja. "IMAGE CAPTION GENERATOR USING CNN AND LSTM." *Turkish Journal of Computer and Mathematics Education* 15, no. 3 (2024): 266-277.
- [18] Poddar, Ayush Kumar, and Rajneesh Rani. "Hybrid architecture using CNN and LSTM for image captioning in Hindi language." *Procedia Computer Science* 218 (2023): 686-696.
- [19] Verma, Akash, Arun Kumar Yadav, Mohit Kumar, and Divakar Yadav. "Automatic image caption generation using deep learning." *Multimedia Tools and Applications* 83, no. 2 (2024): 5309-5325.
- [20] Devi, Ponnaganti Rama, Mannam Thrushanth Deepak, Morampudi Lohitha, M. Surya Chandra Raju, and K. Venkata. "Image Caption Generator Using VGG and LSTM For Visually Impaired." *International Journal of Advances in Engineering and Management* 5, no. 4 (2023): 576-583.

- [21] Panicker, Megha J., Vikas Upadhayay, Gunjan Sethi, and Vrinda Mathur. "Image caption generator." In International Journal of Innovative Technology and Exploring Engineering (IJITEE), vol. 10, no. 3. 2021
- [22] .Agrawal, Vaishnavi, Shariva Dhekane, Neha Tuniya, and Vibha Vyas. "Image caption generator using attention mechanism." In 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), pp. 1-6. IEEE, 2021.
- [23] Agarwal, Lakshita, and Bindu Verma. "From methods to datasets: A survey on Image-Caption Generators." Multimedia Tools and Applications 83, no. 9 (2024): 28077-28123.
- [24] Raypurkar, Manish, Abhishek Supe, Pratik Bhumkar, Pravin Borse, and Shabnam Sayyad. "Deep learning based image caption generator." International Research Journal of Engineering and Technology (IRJET) 8, no. 03 (2021).
- [25] Seo, Paul Hongsuck, Piyush Sharma, Tomer Levinboim, Bohyung Han, and Radu Soricut. "Reinforcing an image caption generator using off-line human feedback." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 03, pp. 2693-2700. 2020.
- [26] Han, Seung-Ho, and Ho-Jin Choi. "Domain-specific image caption generator with semantic ontology." In 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 526-530. IEEE, 2020.
- [27] Wang, Haoran, Yue Zhang, and Xiaosheng Yu. "An overview of image caption generation methods." Computational intelligence and neuroscience 2020, no. 1 (2020): 3062706.
- [28] Shukla, Sujeet Kumar, Saurabh Dubey, Aniket Kumar Pandey, Vineet Mishra, Mayank Awasthi, and Vinay Bhardwaj. "Image caption generator using neural networks." International Journal of Scientific Research in Computer Science, Engineering and Information Technology 7, no. 3 (2021): 1-7.
- [29] Kameswari, A. V. N., and B. Prajna. "Image caption generator using deep learning." Int J Res Appl Sci Eng Technol 9, no. 10 (2021): 1554-1564.
- [30] Anu, Maria, and S. Divya. "Building a voice based image caption generator with deep learning." In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 943-948. IEEE, 2021.
- [31] adityajn105: Flickr 8k dataset, www.kaggle.com/datasets/adityajn105/flickr8k