A Novel Approach for Generating Captions for Visuals Using Deep Learning

Koteswara Rao Velpula[1], Mohan Kalyan Guntupalli[2], Venkatesh Katuri[3], Saidaiah Kandrakunta[4]

1(Assistant Professor, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: koteswararao@vvit.net)
2(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: 20BQ1A0567@vvit.net)
3(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: 20BQ1A0598@vvit.net)
4(UG Student, Department of CSE, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, Email: 20BQ1A0589@vvit.net)

Abstract:
Visual captioning involves creating descriptions of what is happening in an image. It helps build descriptions that explain the content of images. This paper introduces an innovative approach to caption generation using deep learning, specifically utilizing Convolutional Neural Networks (CNNs) for extracting image features and Long Short-Term Memory (LSTM) networks for generating sequences. Additionally, we incorporate nucleus sampling, a probabilistic technique, to improve the diversity and quality of the generated captions, offering more insightful and contextually relevant descriptions for images. This paper marks a significant advancement in automatic image captioning, showcasing the effectiveness of deep learning techniques combined with sophisticated sampling strategies to produce compelling and informative image descriptions.

Keywords: **Convolutional Neural Network(CCN), Long Short-Term Memory(LSTM), Nucleus Sampling, Natural Language Processing(NLP),Feature Extraction.**

I. INTRODUCTION

In recent years, the field of computer vision has undergone significant progress, mainly driven by the rapid advancements in deep learning techniques[8]. A captivating application in this domain is image caption generation, aiming to automatically create descriptive and coherent textual explanations for images. This task intersects computer vision and natural language processing (NLP) and holds immense potential for applications such as assistive technologies, content retrieval, and human-computer interaction.

Traditional approaches to image captioning relied on handcrafted features and rule-based systems[12], struggling to capture intricate semantic relationships between visual and textual information. However, the emergence of deep learning, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), has brought notable advancements. CNNs are proficient in extracting high-level visual features, while RNNs, particularly long short-term memory (LSTM) networks, excel at generating word sequences, making them suitable for sequential data like language. [9]

In this paper, we present a novel approach to image caption generation that leverages the capabilities of deep learning, specifically utilizing CNNs for extracting image features and LSTM networks for generating sequences. To address the challenge of generating diverse and high-quality captions, we

introduce nucleus sampling, a probabilistic technique that guides the model toward producing more varied and contextually relevant captions.

II. RELATED WORK

Numerous studies have delved into the utilization of various captioning systems for image description across diverse domains, driving progress in the field of computer vision and natural language processing.

[1] This system incorporates CNN, GRU (Gated Recurrent Unit), and an Attention mechanism. The inclusion of the Attention mechanism enhances the system's ability to identify specific objects within the images.

[2] The findings from this study reveal that the usage of LSTM (Long Short-Term Memory) as a decoder with CNN produces better results when compared to the CNN-RNN architecture. This is because RNN is not capable of remembering long-term dependencies

[3] This system generates domain-specific captions. The study introduces the usage of an Image Caption Generator with a Caption Reconstructor, where the caption reconstructor modifies the generated caption related to the domain. However, the findings indicate that the performance is not up to the mark.

[4] This work involves a comparison of VGG+GRU and VGG+LSTM models to determine which serves as a better decoder in the caption generation process. BLEU scores are used for the comparison. The findings reveal that LSTM works slightly better than GRU, but GRU takes less time to process.

[5] This work utilizes the RESNET-LSTM architecture for caption generation and finds that it performs more efficiently compared to the CNN-RNN model.

In general, these related studies employed various approaches for generating captions corresponding to the provided images.

III. METHODOLOGY

This system is a combination of Convolutional Neural Network(CNN),Long Short Term Memory(LSTM) and Nucleus sampling techniques.

A. Convolutional Neural Network (CNN)

A Convolutional Neural Network (CNN) is commonly used in caption generators to extract crucial visual features from input images. Its architecture processes raw pixel values, transforming them into a hierarchical representation that captures significant visual features.[8,14]

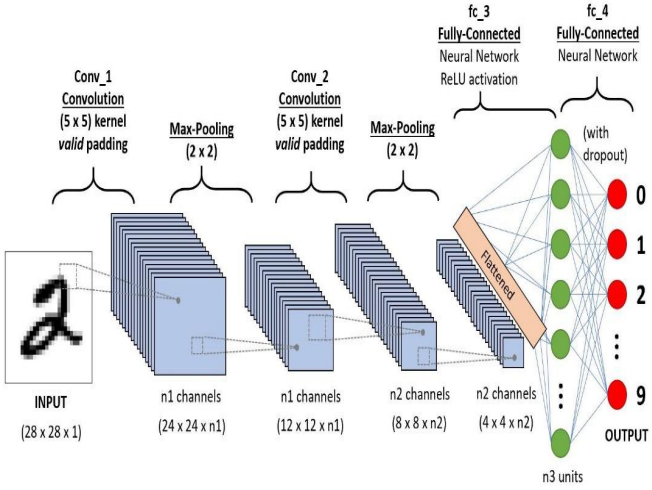

Fig. 1 Convolutional Neural Network(CNN) Architecture

Convolutional layers: These layers use convolutional operations to detect local patterns and features in an image.[8]

Pooling Layers: Pooling layer reduces the spatial size of the feature map and focuses on the most important information.

Flattening Layer: Flattening converts multidimensional feature maps into one-dimensional vectors.

Fully Connected Layers: These layers process the flattened vector and generate high-level features that are relevant to the image.

For this work, Visual Group Geometry(VGG) network is a better model, which is Deep CNN for large scale image recognition. It is available in 16 layers as well as 19 layers.

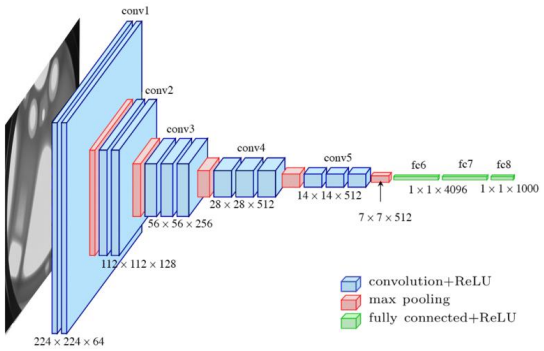

Fig. 2 VGG16 Architecture

B. Long Short Term Memory(LSTM)

Long Short-Term Memory (LSTM) cells, an improvement in Recurrent Neural Networks (RNNs), stand out in preserving information for extended periods. This addresses the vanishing gradient problem that traditional RNNs face, hindering their ability to recall important words in sequence, especially in tasks like caption generation. When generating captions, the LSTM takes in the VGG output (image feature vector) and vocabulary from training captions. The first layer of LSTM uses this information to begin the caption, relying on prior training. Subsequent words are then generated based on the image vector and words already created, resulting in the final caption for the given image.[5,6]

C. Nucleus Sampling

Nucleus sampling is a technique used in natural language processing and machine learning for text generation, also known as top-p sampling or top-k sampling. It is one of the best ways to control the variety and randomness of the generated text. Instead of sampling from the entire probability distribution of words, the model narrows down the selection to the top-p most likely words with the help of this technique..

$$\sum_{x \in V^{(p)}} P(x|x_{1:i-1}) \geq p$$

where p ∈[0,1] is the probability parameter and $V^{(p)}$ is the smallest subset of V.

By combining the strengths of CNN for image feature extraction and LSTM for sequential information processing, augmented with Nucleus Sampling, aims to improve the creativity and diversity of generated captions.

IV. PROPOSED MODEL

In this paper, we propose this model to enhance the creativity and diversity in generating captions for images.

A. Dataset collection

Various datasets, such as ImageNet, COCO, FLICKR 8K, and FLICKR 30K, are available for training deep learning models to generate image captions. This paper specifically utilizes the FLICKR 8K dataset for training the model, as it proves to be effective for the task of training Visual Caption Generating Deep Learning Models. The FLICKR 8K [7] dataset consists of 8000 images.

B. Image preprocessing

Once the datasets are loaded, the next step involves preprocessing the images to make them compatible with the VGG16 model. Since VGG16 requires consistent input sizes, all images need to be resized to a uniform dimension, typically 224X224X3, to pass through the Convolution layer effectively.[10]

C. Text preprocessing

After obtaining captions from the FLICKR text dataset for the corresponding images, it is essential to preprocess them for clarity and ease of vocabulary creation during deep learning model training. This involves removing any numbers present in the captions, eliminating white spaces. To avoid ambiguity, all uppercase letters in the captions are converted to lowercase. For effective training and testing, the model generates captions one word at a time, utilizing previously generated words alongside image features as inputs. To signal the neural network about caption initiation and conclusion, "<st>" and "<end>" tags are added at the beginning and end of caption.[10,11]

D. Defining the model

Having collected and preprocessed the dataset, the next step involves defining the caption generation model. Our proposed approach is the VGG (Visual Geometry Group) - LSTM (Long Short-Term Memory) model. In this configuration, VGG serves as the encoder, extracting image features and converting them into a single-layered vector. These features are then passed as input to the LSTM, which functions as the decoder. To enhance the caption generation diversity, we introduce Nucleus Sampling during the decoding process. The LSTM, acting as the decoder, utilizes Nucleus Sampling to select the most probable words from a subset, ensuring a balance between randomness and coherence. This combination ensures a more nuanced and diverse generation of captions for the given images.

E . Model Training and Optimization

Following the model definition, the training phase involves utilizing the training set to optimize the selected loss function, such as Categorical Cross-Entropy, employing the Adam optimizer. To enhance generalization performance and mitigate overfitting, incorporate techniques like early stopping and dropout regularization. Iteratively fine-tune essential hyperparameters, including learning rate, batch size, and the number of epochs, through continuous training and validation processes. This comprehensive approach ensures the model is effectively trained and capable of generating accurate and diverse captions for images.

F. Model Evaluation

Evaluate the trained model's performance using the BLEU score,[2] which is a widely accepted metric for assessing the quality of generated captions. The BLEU score quantifies the similarity between the model's output and reference captions, providing a comprehensive measure of the model's ability to generate captions. This evaluation helps gauge the overall effectiveness of the visual caption generation model and its capability to produce relevant and accurate descriptions for the given images.

G. Model Architecture

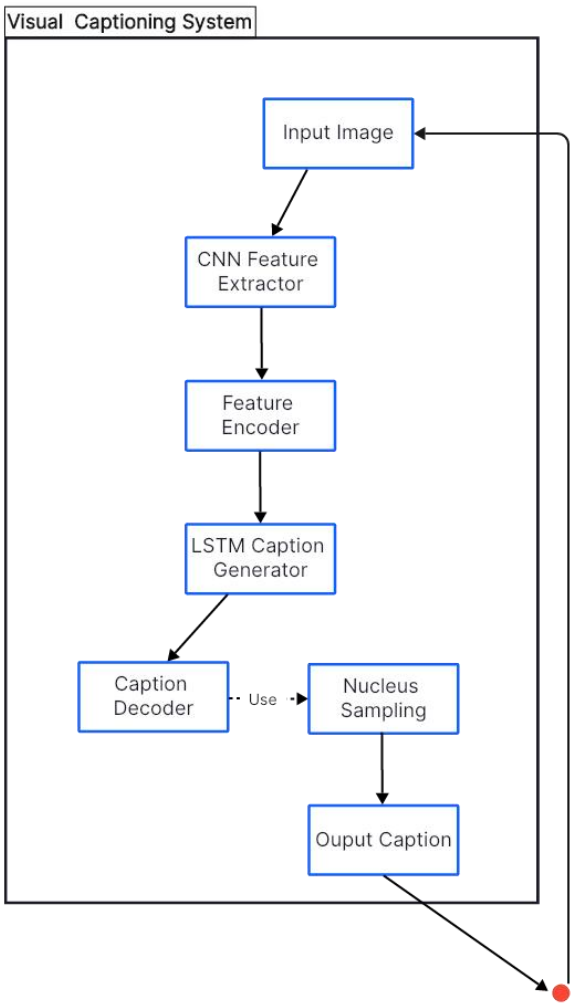

Fig. 3 Process Flow

V. RESULT AND DISCUSSION

After defining and fitting the model, we conducted training for 20 epochs. It was noted that in the initial epochs, the accuracy was considerably low, and the generated captions exhibited less relevance to the provided test images. However, after training for a minimum of 13 epochs, we observed a

noticeable improvement, with the generated captions showing some relevance to the test images.

Furthermore, training the model for the full 20 epochs resulted in reduction of loss to 2.1725 and the generated captions exhibited a higher degree of relevance to the given test images.

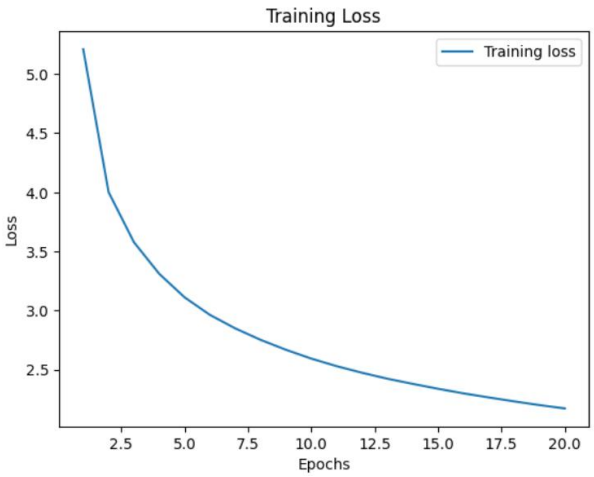

Fig. 4 Loss vs Epoch.

Below are the captions generated by our caption generator.

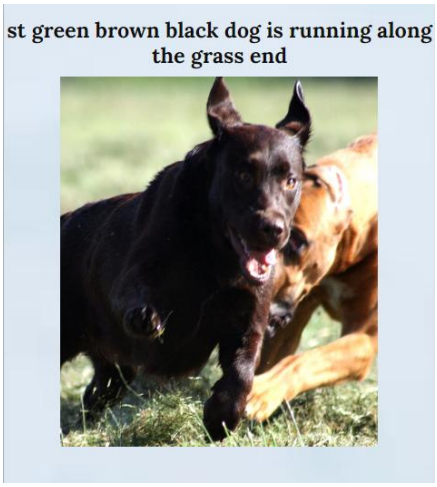

Fig. 5 Generated Caption: green brown black dog is running along the grass.

Fig. 6 Generated Caption: peach little little girl eats peach.

Fig. 7 Generated Caption: baby baby sitting in green chair.

Apparently, the network generated captions are not all perfect, some of which miss important information in the image and others have misidentified visual features. For example, in the figure below, the generated caption is "little blonde girl is running through the grass," but the girl is playing in the mud.

Fig. 8 Generated Caption: little blonde girl is running through the grass.

By using the Flickr8k dataset for training model and running test on the 1000 test images available in dataset results in BLEU score of 0.53356.

TABLE 1

Comparison with other models which used FLICKR 8K dataset

Technique	Epochs	Loss Value	BLEU Score
CNN-LSTM[2]	50	3.74	0.54
CNN-LSTM[13]	30	0.36	0.61
CNN-LSTM-Nucleus Sampling	20	2.1725	053356

VI. CONCLUSION AND FUTURE SCOPE

This paper introduces a visual captioning deep learning model, employing the VGG-LSTM architecture alongside Nucleus sampling to generate captions for provided images. The model is trained using the Flickr 8k dataset, with VGG16 serving as the convolution layer architecture for feature extraction. The extracted image features are then provided to Long Short-Term Memory (LSTM) units acting as the decoder, which utilizes Nucleus Sampling to select the most probable words from a subset, ensuring a balance between randomness and coherence and captions are generated. This visual captioning deep learning model proves to be highly valuable for analyzing large amounts of unstructured and unlabeled data, particularly in guiding self-driving cars and developing software for aiding individuals with visual impairments.

Even though deep learning has progressed a lot, generating the perfect caption is still tricky. This is because of issues like the need for powerful hardware, lack of a proper programming logic, and models that can't generate exact captions as accurately as humans. Looking ahead, as hardware and deep learning models get better, we aim to generate more accurate captions. We also plan to expand this model to create a complete Image-to-Speech conversion, turning image captions into spoken words. This could be really useful for blind people.

REFERENCES

1. *Ansar Hani, Najiba Tagougui & Monji Kherallah "Image Caption Generation Using A Deep Architecture" , ACIT,2019, DOI: 10.1109/ACIT47987.2019.8990998*
2. *Chetan Amritkar,Vaishali Jabade "Image Caption Generation using Deep Learning Technique", ICCUBEA,2018, DOI: 10.1109/ICCUBEA.2018.8697360*
3. *Seung-Ho Han and Ho-Jin Choi "Domain-Specific Image Caption Generator with Semantic Ontology",IEEE International Conference on Big Data and Smart Computing(BigComp),2020, DOI: 10.1109/BigComp48618.2020.00-12*
4. *Grishma Sharma,Priyanka Kalena,Nishi Malde,Aromal Nair,Saurabh Parkar "Visual Image Caption Generator Using Deep Learning", ICAST,2019, DOI: 10.2139/ssrn.3368837*
5. *Aishwarya Maroju, Sneha Sri Doma, Lahari Chandarlapati "Image Caption Generating Deep Learning Model", IJERT,Vol. 10,Issue 9, September 2021, DOI: 10.17577/IJERTV10IS090120*
6. *M.Sailaja,K.Harika,B.Sridhar,RajanSingh,V.Charitha,Koppula Srinivas Rao, "Image Caption Generator using Deep Learning", IEEE,2022, DOI: 10.1109/ASSIC55218.2022.10088345*
7. *Dhirendra Parate, Minu Choudhary "Image Caption Generator using deep learning with Flickr Dataset", IJRTI, Volume 7, Issue 8,2022*
8. *Smriti Sehgal,Jyoti Sharma,Natasha Chaudhary, "Generating Image Captions based on Deep Learning and Natural language Processing", ICRITO, June 2020, DOI: 10.1109/ICRITO48877.2020.9197977*

9. *K. Praveen Kumar, V. Prakash Reddy, G. Indra Karan Reddy, N.S. Ganesh, "Image Caption Generator Using CNN", IJCRT, Volume 9, Issue 6, June 2021*
10. *P. Srinivasa Rao, Thipireddy Pavankumar, Raghu Mukkera, Gopu Hruthik Kiran, Velisala Hariprasad, "IMAGE CAPTION GENERATION USING DEEP LEARNING TECHNIQUE", IRJMETS, Volume 4, Issue 6, June 2022.*
11. *Palak Kabra, Mihir Gharat, Dhiraj Jha, Shailesh Sangle, "Image Caption Generator Using Deep Learning", IJRASET, Volume 10, Issue 10, October 2022, DOI: 10.22214/ijraset.2022.47058*
12. *Tarun Wadhwa, Harleen Virk, Dr. Jagannath Aghav, Savita Borole, "Image Captioning using Deep Learning", IJRASET, Volume 8,Issue 6, June 2020, DOI: 10.22214/ijraset.2020.6232*
13. *A. M. Chandrashekhar, Akash Raj K R, Preetham Jain, Vinayaka Bhat, Nagarjun P R, "Image Captioning using Deep Learning for the Visually Impaired", IJRASET, Volume 9,Issue 7, July 2021, DOI: 10.22214/ijraset.2021.36267*
14. *Anish Banda1, Harshavardhan Manne2, Rohan Garakurthi3 "Image Captioning using CNN and LSTM ", Volume 9, Issue 7, August 2021, DOI: 10.22214/ijraset.2021.37846*