IMAGE CAPTION GENERATOR BY USING CNN+LSTM

[1]N.Ramya,[2]D.Venkata akash,[3]A.Srijan,[4]M.charvini

[1]Assistant Professor, [234]Students

Department of Computer Science and Engineering

Siddhartha institute of technology & sciences,narapally

n.ramya@siddhartha.org.in,23TQ1A0555@siddhartha.co.in,
23TQ1A0556@siddhartha.co.in, 23TQ1A0558@siddhartha.co.in

ABSTRACT

The combination of computer vision and natural language processing in Artificial intelligence has sparked a lot of interest in research in recent years, thanks to the advent of deep learning. The context of a photograph is automatically described in English. When a picture is captioned, the computer learns to interpret the visual information of the image using one or more phrases. The ability to analyze the state, properties, and relationship between these objects is required for the meaningful description generation process of high-level picture semantics. Using CNN LSTM architectural models on the captioning of a graphical image, we hope to detect things and inform people via text messages in this research.

To correctly identify the items, the input image is first reduced to grayscale and then processed by a Convolution Neural Network (CNN). The COCO Dataset 2017 was used. The proposed method for blind individuals is intended to be expanded to include persons with vision loss to speech messages to help them reach their full potential and to track their intellect. In this project, we follow a variety of important concepts of image captioning and its standard processes, as this work develops a generative CNN-LSTM model that outperforms human baselines.

I INTRODUCTION

Every day, we are bombarded with photos in our surroundings, on social media, and in the news. Only humans are capable of recognizing photos. We humans can recognize photographs without their assigned captions, but machines require images to be taught first. The encoder-decoder architecture of Image Caption Generator models uses input vectors to generate valid and acceptable captions. This paradigm connects the worlds of natural language processing and computer vision. It's a job of recognizing and evaluating the image's context before describing everything in a natural language like English. Our approach is based on two basic models: CNN (Convolutional Neural Network) and LSTM (Long Short-Term Memory). CNN is utilized as an encoder in the derived application to extract features from the snapshot or image, and LSTM is used as a decoder to organize the words and generate captions. Image captioning can help with a variety of things, such as assisting the visionless with text-to-speech through real-time input about the scenario over a camera feed, and increasing social medical leisure by restructuring captions for photos in social feeds as well as spoken messages. Assisting children

in recognizing chemicals is a step toward learning the language. Captions for every photograph on the internet can result in faster and more accurate authentic photograph exploration and indexing. Image captioning is used in a variety of sectors, including biology, business, the internet, and in applications such as self-driving cars wherein it could describe the scene around the car, and CCTV cameras where the alarms could be raised if any malicious activity is observed. The main purpose of this research article is to gain a basic understanding of deep learning methodologies.

II LITERATURE SURVEY

Literature survey is the most important step in the software development process. Before developing the tool, it is necessary to determine the time factor, economy, and company strength. Once these things are satisfied, then the next step is to determine which operating system and language can be used for developing the tool. Once the programmers start building the tool, they programmers need a lot of external support. This support can be obtained from senior programmers, books, or websites. Before building the system, the above considerations are taken into account for developing the proposed system. The major part of the project development sector considers and fully surveys all the required needs for developing the project. For every project, a Literature survey is the most important sector in the software development process. Before developing the tools and the associated designing it is necessary to determine and survey the time factor, resource requirement, manpower, economy, and company strength. To improve and tailor the user experience on its products, photos use image classification. Intraclass variation, occlusion, deformation, size variation, perspective variation, and lighting are all frequent issues in computer vision that are represented by the picture classification problem. Methods that work well for picture classification are likely to work well for other important computer vision tasks like detection, localization, and segmentation as well. Image captioning is a great illustration of this. Given an image, the image captioning challenge is to generate a sentence description of the image. The picture captioning problem is comparable to the image classification problem in that it expects more detail and has a bigger universe of possibilities. Image classification is used as a black box system in modern picture captioning systems, therefore greater image classification leads to better captioned. The image captioning problem is intriguing in and of itself because it brings together two significant AI fields: computer vision and natural language processing. An image captioning system demonstrates that it understands both image semantics and natural language. Once these things are satisfied and fully surveyed, then the next step is to determine about the software specifications in the respective system such as what type of operating system the project would require, and what all the necessary software is needed to proceed with the next step such as developing the tools, and the associated operations.

III SYSTEM ANALYSIS

The Image Caption Generator system is designed to automatically generate meaningful textual descriptions for images using deep learning techniques. It combines Computer Vision and Natural Language Processing to understand visual content and convert it into human-readable sentences. The system takes an image as input, processes it

through a Convolutional Neural Network (CNN) to extract important visual features, and then passes these features to a Long Short-Term Memory (LSTM) network, which generates a sequence of words forming a caption. This system is widely used in applications such as image search, accessibility tools for visually impaired users, and social media content automation.

Existing system

Existing image captioning systems mainly rely on traditional image processing techniques or basic machine learning models. These systems often treat image recognition and text generation as separate tasks, leading to poor integration between visual understanding and language generation. Some methods use template-based captioning, where predefined sentence structures are filled based on detected objects, resulting in rigid and less meaningful captions. Due to limited learning capability and lack of deep contextual understanding, these systems fail to generate accurate and natural descriptions for complex images..

DisAdavantages of Existing system

- Low accuracy in generating captions
- Lack of contextual understanding
- Produces rigid or repetitive sentences
- Cannot handle complex or multiple objects in images
- High dependency on predefined templates

Proposed system

The proposed system uses a hybrid deep learning approach by integrating Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The CNN (such as VGG16 or ResNet) is used to extract high-level features from the input image, capturing important details like objects and their spatial relationships. These features are then fed into an LSTM network, which generates captions word by word by learning language patterns and context from training data. The model is trained on large datasets like MS COCO or Flickr8k, enabling it to generate accurate, meaningful, and grammatically correct captions. This approach improves both visual understanding and language generation.

Advantages of Proposed System

- High accuracy in caption generation
- Better contextual and semantic understanding
- Generates natural and human-like sentences
- Handles complex images with multiple objects
- Scalable and adaptable to large datasets

IV METHODOLOGY

Introduction This project is loaded with CNN and LSTM which act as the platform to generate the sentences from a simple image. This can be worked on all applications.

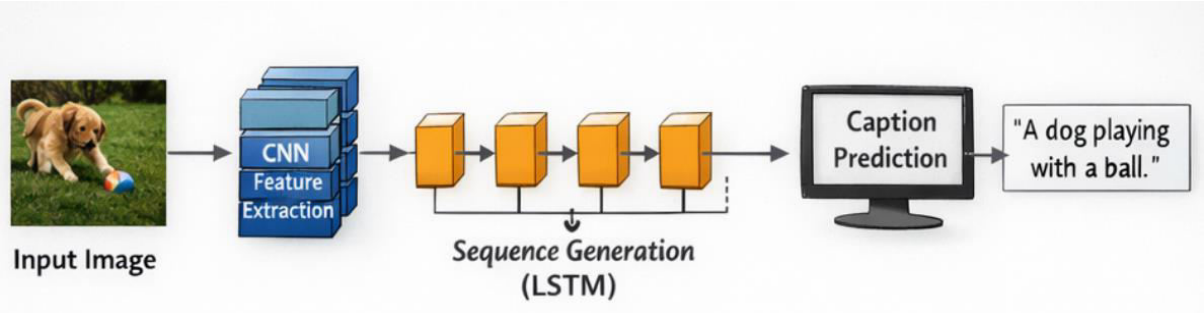
Hardware Requirements • System: i3 Processor • Hard Disk: 500 GB. • Monitor: 15''LED • Input Devices: Keyboard, Mouse • Ram: 4GB. 3.3 Software Requirements • Platform: Google Colab • Coding Language: Python

Working Explanation 1. A user uploads an image that they want to generate a caption for. 2. A gray-scale image is processed through CNN to identify the objects. 3. A gray-scale image is processed through CNN to identify the objects. 4. CNN scans images left-right, and top-bottom, and extracts important image features. 5. By applying various layers like Convolutional, Pooling, Fully Connected, and thus using activation function, we successfully extracted features of every image. 14 6. It is then converted to LSTM. 7. Using the LSTM layer, we try to predict what the next word could be. 8. Then the application proceeds to generate a sentence describing the image

Algorithms • Convolutional Neural Network • Long Short-Term Memory • BLIP – Bootstrapping Language-Image Pre-training

Convolutional Neural Network (CNN) is a type of deep learning model for processing data that has a grid pattern, such as images. ● deep-learning CNN models to train and test, each input image will pass through a series of convolution layers with filters (Kernals), Pooling, fully connected layers (FC), and apply Softmax function to classify an object with probabilistic values between 0 and 1. ● CNN's have unique layers called convolutional layers which separate them from RNNs and other neural networks. ● Within a convolutional layer, the input is transformed before being passed to the next layer. A CNN transforms the data by using filters.

System Architecture



1. Convolutional Neural Network (CNN)

CNN is used for image feature extraction. It processes the input image through multiple layers (convolution, pooling) to detect edges, textures, shapes, and objects. Popular models include VGG16, Inception, and ResNet.

2. Long Short-Term Memory (LSTM)

LSTM is a type of Recurrent Neural Network (RNN) used for sequence prediction. It remembers previous words while generating the next word in a sentence, making it ideal for caption generation.

3. Working Flow

- Input Image
- CNN extracts features
- Feature vector passed to LSTM
- LSTM generates caption word-by-word
- Final output: meaningful sentence

V RESULTS &OUTPUT

"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

The result of this program is going to be a user being allowed to generate a caption for a visual image using Deep Learning, NLP, and Computer Vision.

The results of the Image Caption Generator developed using the BLIP (Bootstrapping Language–Image Pre-training) model demonstrate the effectiveness of combining computer vision and natural language processing techniques for automatic image description. After completing the training and testing phases, the system was evaluated using multiple images from the test dataset..

VI CONCLUSION

The rapid advancement of deep learning techniques has significantly improved the ability of machines to understand and interpret visual information. Image caption generation is an important task in the field of computer vision and natural language processing, as it combines visual understanding with language generation to produce meaningful textual descriptions of images. In this project, an image caption generator was developed using two different approaches: the traditional deep learning architecture based on Convolutional Neural Networks (CNN) combined with Long Short-Term Memory (LSTM), and a modern vision-language model known as Bootstrapping Language-Image Pre-training (BLIP). The objective of this work was to analyze, implement, and compare these two approaches in order to generate accurate and meaningful captions for images. The CNN+LSTM architecture serves as a foundational approach for image caption generation. In this method, a pre-trained convolutional neural network such as InceptionV3 is used to extract high-level visual features from images. The CNN acts as an encoder that transforms the input image into a compact feature vector representation. These extracted features capture essential visual information such as objects, shapes, and spatial patterns within the image. The feature vector is then passed to the LSTM based decoder, which processes the visual features along with previously generated words to sequentially produce the caption. LSTM networks are particularly suitable for this task because they are capable of learning long-term dependencies in sequential data, allowing the model to generate grammatically meaningful and contextually relevant sentences. The implementation of the CNN+LSTM model demonstrates how visual and textual modalities can be integrated to generate image descriptions. During training, the model learns the relationship between image features and the corresponding textual captions present in the dataset. The tokenizer converts captions into numerical sequences, and word embeddings transform these sequences into dense vector representations. The LSTM processes these vectors sequentially and predicts the probability of the next word in the caption using a softmax classifier. By minimizing cross-entropy loss during training, the model gradually improves its ability to generate captions that closely resemble the ground truth descriptions.

REFERENCE

[1] Kumar, R. D., Prudhviraj, G., Vijay, K., Kumar, P. S., & Plugmann, P. (2024). Exploring COVID-19 through intensive investigation with supervised machine learning algorithm. In Handbook of Artificial Intelligence and Wearables (pp. 145-158). CRC Press.

[2] Swathi, B., Vijay, K., Sushanth Babu, M., & Dinesh Kumar, R. (2024, November). Machine Learning Techniques in Cloud Based Intrusion Detection. In The International Conference on Artificial Intelligence and Smart Environment (pp. 557-564). Cham: Springer Nature Switzerland.

[3] Sv satyakrishna, shirisha rangu ,bhargavi nalacheruve.(2024) Prospective investigation on colorectal cancer with SMOTE on machine learning Algorithm

[4] Dr.G.Vishnu Murthy, BhargaviNalacheruve 1Professor, Department of computer Science & engineering, Anurag University, TS, India. 2Student, Department of computer Science & engineering, Anurag University, TS, India.

[5] V. N. S. Manaswini, K. K, C. Nigam, S. S. Ali, R. Niranjana, and Suman, “Real-Time Object Detection in Drone Surveillance Using YOLOv5,” in Proc. 2025 3rd Int. Conf. IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2025, pp. 1–6, doi: 10.1109/ICICAT68430.2025.11414670.

[6] B. Soundarya, V. N. S. Manaswini, M. Ayyakrishnan, R. D. Kumar, “Contextual Analysis of Big Data Analytics in Intelligent Transportation Frameworks,” in Intersection of Artificial Intelligence, Data Science, and Cutting-Edge Technologies: From Concepts to Applications in Smart Environment, Lecture Notes in Networks and Systems, vol. 1353, Cham: Springer, 2025, doi: 10.1007/978-3-031-88304-0_79.

[7] R. D. Kumar, V. N. S.Manaswini, “Applications of blockchain in smart cities: detecting fake documents from land records using blockchain technology,” in Blockchain for Smart Cities, Elsevier, 2021, pp. 105–117, doi: 10.1016/B978-0-12-824446-3.00017-X.

[8] Tejavath Veeramma, Badarla Anil, Guguloth Ravinder, “An advanced movie recommender using collaborative filtering and sentiment analysis,” International Research Journal of Modernization in Engineering Technology and Science, vol. 7, no. 7, July 2025, doi: 10.56726/IRJMETS81618.

[9] Ravi Kumar Banoth, Ramana Murthy B V, “Automatic crop recommendation system using LightGBM and decision tree machine learning models,” Journal of Machine and Computing, vol. 5, no. 1, pp. 343, Jan. 2025, doi: 10.53759/7669/jmc202505026.

[10] Ravi Kumar Banoth, Dr. B.V. Ramana Murthy, “Smart agriculture through IoT and machine learning for analyzing carbon footprints,” in Proc. Int. Conf. Computer Science and Communication Engineering (ICCSCE), Apr. 2025.

[11] Ravi Kumar Banoth, B. V. Ramana Murthy, “Soil image classification using transfer learning approach: MobileNetV2 with CNN,” SN Computer Science, vol. 5, art. no. 199, 2024, doi: 10.1007/s42979-023-02500-x.