Visual Image Caption Generator Using Deep Learning

Md. Faizan Ahmad Zulfiqar Ahmad
B.Voc in Artificial Intelligence and Data Science Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic

Abstract—

The intersection of computer vision and natural language processing has led to significant advancements in artificial intelligence. One such application is automatic image captioning, where a system generates descriptive sentences for images. This research focuses on a generative model combining Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to generate accurate captions for images. Using the COCO 2017 dataset, the proposed CNN-LSTM architecture extracts visual features and translates them into coherent English descriptions. The model aims to assist visually impaired individuals through text-to-speech capabilities and has broader applications in fields such as surveillance, autonomous vehicles, and digital content management

I.INTRODUCTION

Image captioning bridges the gap between computer vision and natural language processing by enabling machines to interpret and describe visual content in natural language. Whilehumans can effortlessly recognize and describe images, machines require sophisticated models to perform this task.

The encoder-decoder architecture, leveraging CNNs as encoders andL STMs as decoders,has emerged as a robust solution for generating meaningful image captions.This research explores the implementation of such a model and evaluates its performance on a large-scale dataset

II. Architecture and Working

III. CNN as Feature Extractor

CNNs are designed to process data with a grid-like topology, such as images. They scan images from left to right and top to bottom, extracting salient features through convolutional and pooling layers.The CNN used in this research is based on the Xception architecture, pre-trained on large image datasets for effective feature extraction.The output is a feature map that encapsulates the essential visual information of the input image

LSTM as Sequence Generator

LSTMs, a type of Recurrent Neural Network (RNN), are adept at handling sequential data and overcoming the vanishing gradient problem inherent in traditional RNNs. In this model, the LSTM receives the feature map from the CNN and generates a sequence of words, forming a descriptive caption. The LSTM's ability to retain information over long sequences is crucial for generating coherent and contextually relevant sentences

IV. METHODOLOG

Dataset:COCO2017dataset,which contains over 120,000 images with 5 captions each.

- Preprocessing: Images resized and normalized; captions tokenized and padded.
- FeatureExtraction:CNNmodel(Xception)used to extract image embeddings.
- Model Architecture: Combined CNN encoder and LSTM decoder to generate captions.
- Training:Trainedwithcategoricalcross-entropy loss and Adam optimizer.

ModelWorkflow

Input image is pre-processed and converted to grayscale.CNNextractsfeaturesandproducesafeature map.
LSTMreceivesthefeaturemapandgeneratesacaptionwor d byword.
The process continues until an end-of-sequence token isgenerated

V. DATASETS

TheCOCO2017dataset,comprisingimages categorized into 12 main types with 80 subcategories, wasusedfortrainingandevaluation.Eachimageis annotated with five different captions, providing a rich source for learning diverse descriptions.The dataset was pre-processed to ensure compatibility with the model architecture

VI. RESULTSANDDISCUSSION

The model was trained over 30 epochs, with a total of 130 iterations. Performance was assessed using standard metrics such as accuracy and confusion matrices. The CNN-LSTM model demonstrated improved accuracy over baseline methods, effectively generating captions that closely matched human-generated descriptions. The attention mechanism further enhanced the model's ability to focus on relevant image regions, resulting in more precisecaptions

VII. APPLICATIONSANDFUTURESCOPE

Aiding visually impaired individuals via text-to-speech. Surveillance enhancement by describing CCTV footage.

- Automatic content tagging in socialmedia.
- Assisting children in educational learning.

Future improvements include:

- Integration with attention mechanism for better context handling.
- Use of transformer-based architectures.
- Multilingual caption generation.
- Real-time implementation on IoT devices using edge computing.

VIII. CONCLUSION

Future work will focus on expanding the dataset, incorporating multilingual capabilities,and deploying

This research successfully implements a CNN-LSTM

based image captioning model capable of generating accurate and meaningful descriptions for a wide range of images. The model's effectiveness is evident in its performance on the COCO dataset and its potential applications for visually impaired individuals, surveillance, and content indexing.

the model in real-world scenarios using IoT devices

ACKNOWLEDGMENT

I sincerely thank the faculty of Anjuman-I-Islam Abdul Razzak Kalsekar Polytechnic,especially the Department of Artificial Intelligence and Data Science, for their valuable guidance,continuous support,and encouragement through out this project.

My heartfelt appreciation goes to the Hugging Face open- source community for providing powerful and accessible NLP tools,which were instrumental inimplementing Visual Image Caption Generator Using Deep Learning.

I am also grateful to my peers,mentors,and everyone who directly or indirectly contributed to the successful completion of this work.

REFERENCE

Poddar, Ayush Kumar, and Dr. Rajneesh Rani. "Hybrid Architecture usingCNNandLSTMforImageCaptioninginHindiLanguage." ScienceDirect,2023.

Ghandi, Taraneh, Hamidreza Pourreza, and Hamidreza Mahyar. "Deep Learning Approaches on Image Captioning: A Review." arXiv:2201.12944 [cs.CV], 2022.

"MS-COCO Flicker8k dataset for image captioning."Packt Publishing,2021