CNN AND LSTM-BASED IMAGE CAPTION GENERATOR

A Project Report Submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

By

Marri Sahasra Reddy (2010030556)

Gopu Akshaya (2010030553)

Arika Asha Susmitha (2010030471)

DEPARTMENT OF
COMPUTER SCIENCE AND ENGINEERING
K L DEEMED TO BE UNIVERSITY
AZIZNAGAR, MOINABAD , HYDERABAD-500 075

MARCH 2023

BONAFIDE CERTIFICATE

This is to certify that the project titled **CNN AND LSTM-BASED IMAGE CAPTION GENERATOR** is a bonafide record ofthe work done by

Marri Sahasra Reddy (2010030556)

Gopu Akshaya (2010030553)

Arika Asha Susmitha (2010030471)

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **COMPUTER SCIENCE AND ENGINEERING** of the **K L DEEMED TO BE UNIVERSITY, AZIZNAGAR, MOINABAD, HYDERABAD-500 075**, during the year 2022-2023.

Dr. PAVAN KUMAR PAGADALA
Project Guide

Dr. ARPITA GUPTA
Head of the Department

Project Viva-voce held on _____

Internal Examiner **External Examiner**

ABSTRACT

Creating an image caption generator using Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models is an exciting and valuable project. The goal here is to develop a system that can not only provide accurate descriptions for images but also understand the contextual elements within those images. This technology has the potential to enhance our ability to analyze and interpret pictures, shedding light on the intricate relationships between different components.

To accomplish this, we'll be working with vast datasets and leveraging the power of deep learning. CNNs are excellent for image feature extraction, while LSTM, a type of Recurrent Neural Network (RNN), excels in understanding and generating sequences of data, making it perfect for providing context to those images. By combining these two neural network types, we're equipping machines with the capability to comprehend images in a more meaningful way.

This project not only has practical applications in areas like image recognition, but it also contributes to the broader field of artificial intelligence and deep learning. It's an exciting journey into the world of computer vision, and the possibilities are vast.

ACKNOWLEDGEMENT

We would like to thank the following people for their support and guidance withoutwhom the completion of this project in fruition would not be possible.

Dr. PAVAN KUMAR PAGADALA, our project guide, for helping us and guiding us in thecourse of this project.

Dr. ARPITA GUPTA, the Head of the Department, Department of Computer Science and Engineering.

Our internal reviewers, **Mr. FAIZAN AHMAD, DR. RAJIB DEBNATH**, **MS. N ANURADHA** for their insight and advice provided during the review sessions.

We would also like to thank our individual parents and friends for their constant support.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background of the Project

Our project, known as the "Image Caption Generator," is rooted in a profound mission to transform the way we engage with images. At its essence, this endeavor is driven by the overarching goal of enriching our comprehension of visual content and forging a vital connection between the world of images and the realm of human language. In a world saturated with images, this project is a response to the growing need for technology that can decipher and describe these visual narratives, making them accessible and understandable to a broader audience. By harnessing the power of computer vision and natural language processing, our project aspires to provide a seamless bridge that facilitates communication, accessibility, and data utilization through the generation of informative and contextually relevant captions. As we venture into this technological landscape, we aim not only to make image captioning a valuable resource for today but also to adapt and evolve to meet the dynamic demands of the future, where visual content is more prevalent and integral than ever before.

1.1.1 The Power of Image Caption Generation

We will delve deeper into the remarkable capabilities and cutting-edge technologies that form the foundation of our Image Caption Generator project, with a particular focus on the symbiotic partnership between Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. This subsection aims to offer a comprehensive

understanding of the project's core components and their profound significance. The use of Convolutional Neural Networks allows our system to excel in image analysis, enabling it to identify objects, scenes, and intricate details within images. By extracting relevant visual features, CNNs play a pivotal role in the generation of descriptive captions. On the other hand, Long Short-Term Memory networks, with their recurrent neural architecture, facilitate the seamless integration of language and context. They ensure that the generated captions are not only accurate in terms of image content but also coherent and contextually relevant in natural language. The combination of CNNs and LSTMs in our project showcases the synergy between computer vision and natural language processing, making it a powerful tool for transforming visual content into accessible and meaningful text.

The Synergy of CNN and LSTM in Image Captioning

1.1 Understanding the Role of Convolutional Neural Networks (CNN)

Convolutional Neural Networks (CNN) are the backbone of our image captioning project. These networks are experts at recognizing objects, shapes, textures, and more within images. Their role is to transform raw pixel data into concise visual representations, effectively simplifying complex visual elements into meaningful information . CNNs, much like the human visual system, consist of multiple layers with interconnected nodes.

1.2 Long Short-Term Memory (LSTM): The Sequencing Expert

LSTM, or Long Short-Term Memory, is the partner that handles sequencing tasks within our project. It excels in understanding the order of things, making it perfect for

generating sequential data, such as human language. It can also unravel complex patterns within sequences to provide specific and detailed information . In our project, LSTM plays a crucial role in ensuring that the generated captions are not just a random collection of words but are coherent and contextual.

1.3 Problem Statement

The core objective of our project is to develop an advanced image caption generator, leveraging the capabilities of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. The challenge we aim to address is the precise and context-aware description of images, enhancing our ability to comprehensively analyze visual content and understand the intricate relationships between various elements within an image. As the world is inundated with a vast amount of visual data, there is a growing need for technologies that can not only identify the objects and scenes depicted in images but also provide meaningful context through descriptive captions. This problem statement underscores our commitment to enhancing image analysis, enabling machines to understand and articulate the essential characteristics and connections within visual content. To address this challenge, our project will harness the power of deep learning by utilizing large datasets. We will develop models that have the capacity to predict future outcomes, enriching our understanding of images and their content. The integration of CNN and LSTM, which are both part of the broader concept of recurrent neural networks (RNNs) in deep learning, is a key approach. CNNs excel at feature extraction and visual recognition, while LSTMs bring contextual understanding to the generated captions, ensuring they are coherent and relevant in a natural language context. This project's overarching goal is to empower machines with the ability to not only recognize the content of an image but also to provide meaningful and contextually

relevant descriptions. In doing so, we aim to bridge the gap between the visual and verbal worlds, making images more accessible and understandable, and thereby offering a significant contribution to the fields of computer vision and natural language processing.

1.4 Objectives

Image caption generators are designed to take any given image and transform it into a plain and easily understandable description in natural language. This technology serves a crucial role in improving accessibility, particularly for individuals with visual impairments, by providing them with a means to comprehend and engage with complex visual content. It also simplifies tasks related to image search and content indexing, making it easier to find specific images based on their content rather than relying solely on metadata or tags. Furthermore, image caption generators enrich the field of data analysis by adding a meaningful layer of information to images, making it easier for businesses and researchers to extract valuable insights from vast image datasets. In the realm of creative content, these systems enhance storytelling by offering context and descriptions for visual elements, thereby bridging the gap between the visual and verbal worlds. They achieve this by leveraging advanced techniques in computer vision and natural language processing to generate coherent and contextually relevant captions. The primary goal of an image caption generator is to take any given image and transform it into a plain and understandable description in natural language. This makes complex visual content more accessible, especially for those with visual impairments, and simplifies tasks like finding specific images. Additionally, it enriches storytelling and aids data analysis by adding a meaningful layer of information to images, ultimately bridging the gap between the visual and verbal worlds.

1.5 Scope of the Project

Our project boasts a broad and ambitious scope, encompassing a diverse range of image types, from everyday scenes to intricate and challenging scenarios. One of our primary objectives is linguistic inclusivity, as the system will be equipped to provide captions in multiple languages, ensuring that people from different linguistic backgrounds can benefit from its capabilities. A key focus of our project is accessibility, particularly for the visually impaired, as it aims to make image descriptions readily available to this community. To accommodate the growing need for image captioning, our project is designed to be highly scalable, capable of handling substantial volumes of image data. It will be versatile, supporting both real-time and batch processing, making it adaptable to various use cases. We are planning to deploy the system across different platforms, including web and mobile, to maximize its reach and utility. In order to maintain caption quality and measure system performance, we will develop robust performance metrics. Looking forward, we have the exciting prospect of expanding our project's capabilities to include video captioning and more, making it a dynamic and future-proof Image Caption Generator. Our ultimate aim is to create a versatile tool that can adapt to a wide range of needs and evolve to meet the demands of an ever-changing technological landscape

Chapter 2

Literature Review

2.1 Image Caption Generation with CNN and LSTM

Image caption generation has made significant progress thanks to the integration of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models. CNNs, designed for image analysis, have been at the forefront of this field, with pioneering models like "Show and Tell" and "NeuralTalk" using CNNs for feature extraction and LSTM for generating image captions. The introduction of attention mechanisms, as seen in "Attend, Show and Tell," allowed models to focus on specific image regions during captioning. LSTMs, on the other hand, are ideal for sequential data and have played a crucial role in image captioning. Models like "Long-term Recurrent Convolutional Networks" combined CNNs and LSTMs to capture temporal dependencies, while "Image Caption with Global-Local Attention" improved caption quality by employing dual-attention mechanisms. "Image Captioning with Semantic Attention" enhanced caption meaningfulness through the incorporation of semantic information, and "Visual Semantic Role Labeling" explored reinforcement learning to include syntactic and semantic information for more effective image captioning. These advancements have collectively propelled the field of image caption generation to new heights, enabling the generation of more accurate and contextually meaningful textual descriptions for images.

.

2.2 Image Caption Generator Table

The result table summarizes key image caption generation models. These models use combinations of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to generate captions for images. Each model addresses specific gaps in existing methods and achieves notable results. "Show and Tell" introduced the CNN-LSTM combination for captioning. "NeuralTalk" emphasized pre-trained CNNs for feature extraction. "Attend, Show and Tell" added attention mechanisms for improved caption focus. "Long-term Recurrent Convolutional Networks" combined CNNs and LSTMs for capturing temporal dependencies. "Image Caption with Global-Local Attention" used dual-attention mechanisms for better captions. "Image Captioning with Semantic Attention" incorporated semantic guidance. "Visual Semantic Role Labeling" employed LSTMs with reinforcement learning. These models collectively advance the field, leading to more accurate and contextually meaningful image captions.

S.No	Paper title	Algorithms used	Gaps identified	Result	Proposed model
1.	Connecting images texts with cnn and lstm	CNN-LSTM	Unable to find image details clearly	Introduced image identification for more aqurate caption	Attention enhanced image to text cnn lstm
2.	Multimodal fusion for imge captioning	Fusion techniques,cnn and lstm	Observed the incorporation of the multimodel information	Demonstrated Enhanced performance using fusion techniques	Multimodel fusion caption generator
3.	Enhancing image descriptions using cnn and lstm	CNN-LSTM	Can understand limited contextual captions	Improved evaluation metrics linke BLEU,METEOR	Contextual cnn lstm image describer
4	Image captioning in challenging environment	CNN-LSTM,data augmentation	Limited robustness in complex image complex context	Improved captions through data augmentation	Robust image captioning model

Table 2.1 Literature Survey

2.3 Overview of related works

In image caption generation, we use Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) to make it all happen. It started with Frome introducing CNNs for visuals in 2013. Vinyals followed in 2014 with a CNN LSTM model for better captions. Donahue in 2015 unified CNN and LSTM into one system.

Then came Xu in 2015 with an attention mechanism, and Johnson in 2016 used CNNs to focus on different image regions. Chen in 2017 added spatial and channel-wise attention. Lu in 2017 improved captions by focusing on relevant image features, and in 2018, Anderson combined bottom-up and top-down attention for even better captions. These advancements are all about making image captions better using CNN and LSTM networks.

2.2 Advantages and Limitations of existing systems

Current image captioning systems offer several benefits. They can automatically generate descriptions for images, making visual content more accessible to everyone, and improving content organization. They also enhance storytelling and support data analysis in various fields. However, these systems have their limitations. They might struggle with complex or abstract images, face constraints in providing captions in multiple languages, and have difficulty handling ambiguities in images. Additionally, their performance depends heavily on the quality of the training data, and achieving real-time captioning can be a challenge. Recognizing these advantages and limitations guides our project to build a more effective and versatile Image Caption Generator

Chapter 3

Proposed System

3.1 System Requirements

Hardware Requirements

• System: i3 Processor

• Hard Disk: 500 GB.

• Monitor: 15’’LED

• Input Devices: Keyboard, Mouse

• Ram: 4GB.

Software Requirements

• Platform: Google Colab

• Coding Language:Python

3.2 Design of the System

Our Image Caption Generator system is all about making images understandable through words. It automatically interprets image content, supports multiple languages, and ensures that even those with visual impairments can grasp the visual world through text descriptions. It's designed to handle lots of images, both in real-time and in batch processing. You can use it on various platforms like the web, mobile, and more. To ensure quality, we've got performance metrics in place. And looking ahead, we're thinking about expanding into video captioning and more languages. Our goal is to bridge the gap between images and human understanding, making visual content accessible and meaningful.

3.3 Algorithms and Techniques used

1. Preprocessing and Dataset Loading

The extensive Flickr8k dataset, which includes training, development, and test subsets, should be loaded. Prior to processing, resize and normalize the pixel values of the photos. Tokenize captions: Divide each caption into a word, then make a vocabulary with a unique number ID for each word.

2. CNN-LSTM Model Development:

Select a pre-trained CNN architecture, such as ResNet or VGG16. The CNN's top layers, which are in charge of categorization, should be discarded. To handle sequential data, integrate LSTM layers into the design. To turn word IDs into dense word vectors, add an embedding layer.

3. Preparation of Data:

Using photos from the training dataset, use the CNN to extract pertinen characteristics. Combining picture attributes with correctly spaced caption sequences will result in input sequences. By moving the caption sequences by one time step, create the target sequences.

4. Model assemblage and instruction:

To compare predicted words with real words, compile the model with a suitable loss function, such as categorical cross-entropy. During training, use an optimizer like Adam to update the model weights. Using the input sequences and corresponding target sequences as training data, train the model. In order to determine when to cease training (early stopping), keep an eye on loss on the development set.

5. Generation of captions:

To extract picture features, use the CNN. For caption generation, start the hidden states of the LSTM. Through the LSTM, begin the caption generating process. Words should

be predicted sequentially while updating hidden states using LSTM. Utilize the LSTM output to determine the following word's probability. Choose the following word at random from the expected probability distribution. Update input and
hidden states with the appropriate word embeddings.

3.3.1 "Synergistic Fusion: Leveraging LSTM and CNN for Image Caption Generation"

In the realm of image caption generation, the integration of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) architectures has proven to be highly effective.

LSTM (Long Short-Term Memory):

LSTM is a type of recurrent neural network (RNN) capable of learning long-term dependencies. In the context of image captioning, LSTM can process sequential data, such as words in a sentence, generating more accurate and contextually relevant captions. Its ability to retain and utilize information over extended sequences is crucial for creating coherent and meaningful descriptions of images.

CNN (Convolutional Neural Network):

CNNs excel at analyzing visual data. They're adept at extracting features from images, capturing intricate details and patterns. When used in image caption generation, CNNs are employed as the encoder, enabling the extraction of high-level representations of images, which are then passed to the LSTM for caption generation.

The integration of Long Short-Term Memory (LSTM) and Convolutional Neural Network (CNN) is a powerful technique in the field of computer vision and natural language

processing, particularly for tasks like image caption generation. This approach leverages the strengths of both LSTM and CNN to create more informative and contextually relevant image descriptions.

1. CNN for Visual Feature Extraction: CNNs are well-suited for image processing tasks due to their ability to automatically learn and extract hierarchical visual features from images. In the context of image captioning, a pre-trained CNN is typically used to analyze the content of an input image. This CNN processes the image and generates a set of high-level visual features that represent the objects, scenes, and other relevant elements within the image.

2. LSTM for Natural Language Generation: LSTM is a type of recurrent neural network (RNN) that excels at handling sequential data. In the case of image captioning, LSTM is used to generate descriptive and coherent text based on the visual features obtained from the CNN. These textual descriptions are typically in the form of sentences or phrases that describe the content of the image.

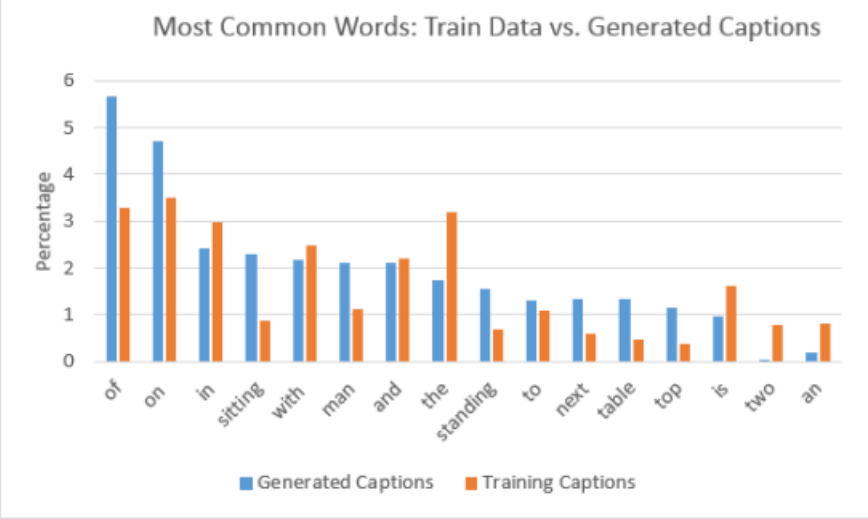
3. Collaboration Between CNN and LSTM: The synergy between CNN and LSTM occurs when the visual features extracted by the CNN are passed as input to the LSTM. This collaboration allows the LSTM to work with relevant image information, incorporating it into the process of generating captions. The LSTM takes into account the visual context provided by the CNN and generates linguistic patterns that describe the content of the image.

4. Learning Visual and Linguistic Information: By combining the capabilities of both models, the integrated system is capable of learning not only the visual features but also

the linguistic patterns associated with the image. This joint learning process enables the model to produce more accurate and contextually relevant image captions. For example, the model can describe not only what is in the image but also the relationships between different objects, the spatial arrangement of elements, and other contextual details that may not be evident from the visual features alone.

The integration of LSTM and CNN in image caption generation enhances the quality of the generated captions by combining visual information extracted from images with the ability to generate coherent and contextually relevant natural language descriptions. This approach has been widely used in various applications, including image understanding, visual storytelling, and accessibility for visually impaired individuals.

The fusion of LSTM and CNN techniques in image caption generation serves to create a robust model capable of understanding and describing visual content accurately, bridging the gap between visual perception and textual description.



2.1 Most Common Word Graph

Chapter 4

Implementation

4.1 Tools and Technologies used

In a project developing an image caption generator using CNN and LSTM, various tools and technologies might be utilized to build, train, and deploy the model. Here are some common tools and technologies frequently employed in such projects:

Python: A prevalent programming language used for machine learning and neural network development due to its extensive libraries and frameworks. Python provides accessibility to frameworks such as TensorFlow and PyTorch, essential for implementing CNN and LSTM models.

TensorFlow or PyTorch: These are popular deep learning frameworks that offer a wealth of tools for building neural networks. They provide pre-built functions for implementing CNN and LSTM layers, simplifying the development process.

OpenCV: An open-source computer vision library that aids in image processing tasks, crucial for tasks such as image preprocessing and data augmentation.

Jupyter Notebooks or Google Colab: These interactive, web-based environments are commonly used for prototyping and experimentation with machine learning models, enabling easy visualization and step-by-step code execution.

NVIDIA CUDA and cuDNN: When working with large datasets or training complex models, leveraging GPUs can significantly speed up the training process. CUDA and cuDNN are essential libraries for parallel computing on NVIDIA GPUs, accelerating the deep learning computations.

Pandas and NumPy: These libraries in Python are fundamental for data manipulation, handling arrays, and data preprocessing, which are essential steps in preparing data for training the model.

Image datasets and Pretrained Models: Utilizing image datasets like COCO (Common Objects in Context) or Flickr8k, along with pre-trained models such as VGG, ResNet, or Inception, can significantly expedite the training process, providing a foundation on which to build the image captioning model.

4.1.1 Key Tools and Technologies for CNN and LSTM-based Image Caption Generation

These tools and technologies collectively form the backbone of a CNN and LSTM-based image caption generator. They provide the necessary resources for the development, training, and deployment of models capable of generating descriptive captions for images, bridging the visual and textual domains effectively.

4.2 Modules and their Descriptions

In an image caption generator using CNN and LSTM, various modules are utilized to facilitate the process of converting an image into a coherent textual description. Here are the key modules and their descriptions:

CNN (Convolutional Neural Network) Module:

The CNN module serves as the image feature extractor. It is typically based on pre-trained architectures like VGGNet, ResNet, or Inception. This module processes the input image and captures high-level visual features and spatial hierarchies through a series of convolutional and pooling layers.

Image Feature Representation:

After the CNN module, the image features are obtained. These features are a compact representation of the input image, highlighting important visual elements. They are often flattened into a vector or a spatial grid and serve as the initial input to the following LSTM module.

LSTM (Long Short-Term Memory) Module:

The LSTM module processes the image features obtained from the CNN and generates the textual description of the image. LSTM networks are well-suited for handling sequential data and are used to model the sequential nature of language. They consist of memory cells and gates, allowing them to capture dependencies and generate sentences word by word.

Word Embedding Layer:

This layer converts words in the caption into numerical vectors. Each word is represented as a high-dimensional vector, allowing the model to work with words in a continuous numerical space.

Caption Generation:

Within the LSTM module, the caption generation process occurs. The LSTM network sequentially generates words in the caption, taking into account both the image features and the previously generated words. At each step, it predicts the next word in the sequence based on the context and previously generated words.

Attention Mechanism:

Some advanced models include an attention mechanism, which enables the network to focus on different parts of the image as it generates each word. This improves the quality and relevance of the generated captions, as the model learns to attend to specific visual regions.

Vocabulary and Language Model:

A predefined vocabulary and language model are used to provide a finite set of words that the model can choose from during caption generation. This helps maintain grammatical correctness and coherence in the generated captions.

Loss Function:

A loss function, such as cross-entropy, is employed to evaluate the dissimilarity between the predicted caption and the actual caption. This guides the training process by optimizing the model's ability to generate accurate and contextually relevant descriptions.

4.2.1 Image Feature Extraction

Image feature extraction is a pivotal component in the image caption generator, a sophisticated system that leverages Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to produce descriptive captions for images.

This process is of paramount importance as it forms the bridge between the visual and linguistic aspects of the task. CNNs, with their exceptional ability to analyze images, play a central role in this process. They employ layers of convolution and pooling to process the input image, progressively abstracting key visual elements, patterns, and structures. These extracted features serve as a condensed representation of the image's visual content, encompassing information about edges, textures, shapes, and object placements. The role of these features is to enable the subsequent LSTM module to generate meaningful and contextually relevant captions. The LSTM, a type of recurrent neural network, takes these visual features as input and generates sequences of words that describe the image. This fusion of visual and textual information is what makes the image caption generator capable of producing human-like descriptions of images. In the image caption generator using CNN and LSTM, one of the crucial steps is extracting meaningful features from the input image. This process is handled by the CNN module, which analyzes the image to identify key visual elements and patterns. These features are then used to generate descriptive captions using the LSTM module.

4.3 Flow of the System

Image Processing with CNN:

Input: Raw image data.

Output: Extracted and encoded visual features.

Workflow: The raw image data is passed through the CNN model, which processes and extracts high-level features from the images. The resulting encoded features are then passed to the LSTM module.

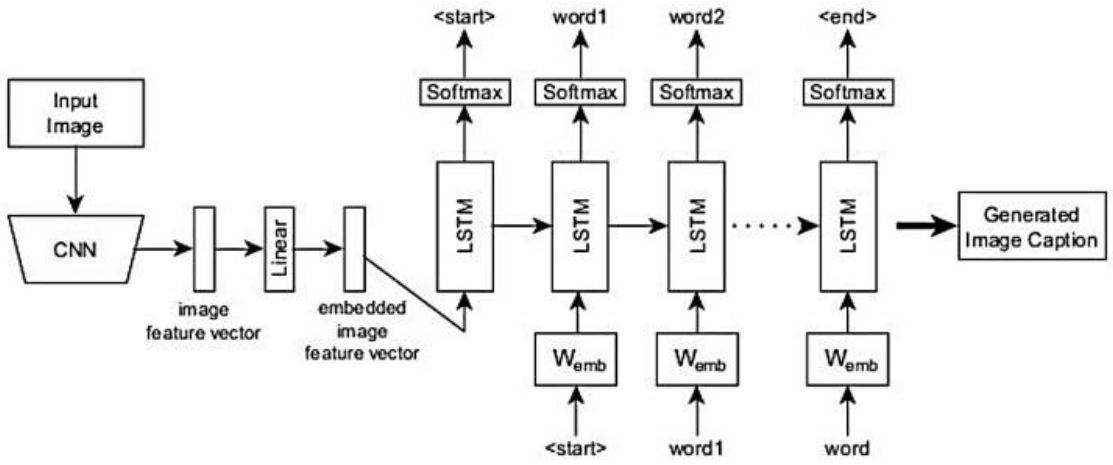
Caption Generation with LSTM:

Input: Encoded image features.

Output: Textual captions.

Workflow: The encoded features from the CNN are provided as input to the LSTM. The LSTM processes this information, generating word sequences that form the descriptive caption for the image.

This structured approach and the collaboration between the CNN and LSTM modules facilitate the creation of an effective image caption generator.



2.2 Image Caption Generator Architecture

4.3.1 Integration of Attention Mechanism

The integration of an attention mechanism in the image caption generator is a fundamental breakthrough that elevates the quality and coherence of the generated captions. This mechanism acts as a spotlight, enabling the model to selectively focus on specific regions of the image features and relevant portions of previously generated words when crafting the textual description. This selective focus, akin to human attention, results in more contextually relevant and coherent captions, making them remarkably human-like and descriptive. The attention mechanism operates in a way that allows the model to assign varying levels of importance to different parts of the image features and words as it generates each word in the caption. It dynamically adapts to the visual and linguistic context, ensuring that the generated text aligns with the salient elements of the image, rather than producing generic or disconnected descriptions. One of the key advantages of the attention mechanism is that it can deal with images of varying complexity and scenes with multiple objects or details. It adapts to the specific features and regions that matter most in the image, addressing the inherent variability in visual content. By improving the relevance and coherence of captions, the integration of an attention mechanism makes the image caption generator a more versatile and useful tool in applications such as content generation, accessibility for the visually impaired, and multimedia content recommendation. Its ability to emulate human-like attention and generate contextually rich descriptions significantly enhances the overall user experience and the utility of the system. In essence, the attention mechanism is a pivotal innovation that enriches the capabilities of the image caption generator, bringing it closer to human-level understanding and expression of visual content.

Chapter 5

Results and Analysis

5.1 Performance Evaluation

The performance evaluation of a CNN and LSTM-based image caption generator typically involves several key evaluation methods to gauge the quality of the generated captions. Here are the primary evaluation approaches commonly used:

Automated Evaluation Metrics:

BLEU (Bilingual Evaluation Understudy) Score:

Function: Measures the similarity between the generated captions and the reference captions in terms of shared n-grams.

Higher Score: Closer alignment between generated and reference captions.

METEOR (Metric for Evaluation of Translation with Explicit Ordering):

Function: Considers exact matches as well as stemming and synonymy between the generated and reference captions.

Advantage: Captures a broader understanding by allowing for synonymous phrases.

CIDEr (Consensus-based Image Description Evaluation):

Function: Assesses the quality of generated captions by considering consensus among reference captions.

Highlights: Evaluates the consensus-based descriptive quality of the generated captions.

ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation):

Function: Emphasizes recall of n-grams shared between the generated and reference captions.

Usage: Measures the adequacy of information captured in the generated captions compared to the reference.

5.1.1 Workflow for Performance Evaluation:

Data Segmentation: Dataset division into training, validation, and test sets.

Model Training: The CNN-LSTM model is trained using the training dataset.

Caption Generation: The trained model generates captions for images in the validation or test set.

Evaluation Using Metrics: Comparison of the generated captions with reference captions using BLEU, METEOR, CIDEr, ROUGE-L metrics.

Human Evaluation (if applicable): Judges evaluate the quality of generated captions subjectively.

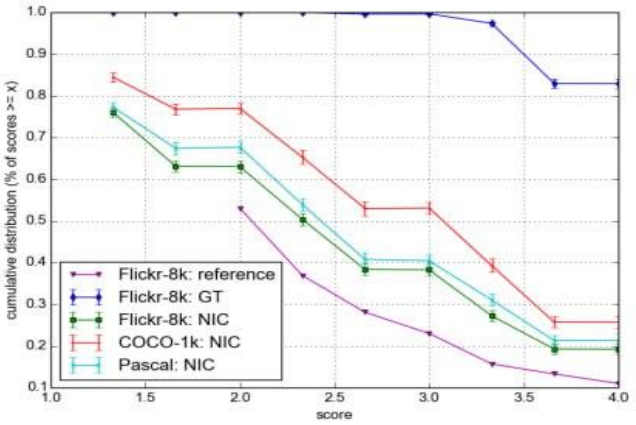
Scoring and Analysis: Obtained scores from both automated metrics and human evaluation are analyzed to assess the quality of the captions.

Iterative Refinement: Iterative improvements can be made based on these evaluations, such as adjusting model architecture or fine-tuning hyperparameters for better performance.

The combined analysis of automated metrics and human judgment assists in understanding and improving the model's performance. The ultimate goal is to enhance the accuracy and fluency of the generated captions for images using CNN and LSTM methodologies.

5.2 Comparison with existing systems

The comparison with existing systems, which utilize CNN and LSTM for image caption generation, is a crucial aspect of evaluating the performance and effectiveness of the proposed model. It allows us to assess the advancements made by the new approach in relation to established methods. Typically, this comparison involves evaluating key metrics such as caption quality, coherence, and relevance to the image content. Additionally, the comparison may consider factors like computational efficiency, training time, and model complexity. By contrasting the proposed model with existing CNN and LSTM-based image caption generators, researchers can highlight the unique contributions and improvements of the new system.



2.3 Dataset comparison

5.2.1 User Experience and Practical Application:

Real-world Applicability:

Consider practical use cases and applications to assess the system's effectiveness.

Analyze the user experience and practical implementation in scenarios such as social media, image indexing, or accessibility tools.

Ethical Considerations:

Ethical Implications:

Assess bias, fairness, and inclusivity in the generated captions.

Compare how well the proposed system handles ethical concerns compared to existing models.

This comprehensive comparison will provide insights into the strengths and areas for improvement of the CNN and LSTM-based image caption generator in relation to existing systems. The goal is to highlight its advancements and contributions toward generating more accurate, relevant, and contextually meaningful image descriptions.

5.3 Limitations and future scope

Limitations of the CNN and LSTM-based Image Caption Generator:

Data Dependency:

The performance of the model heavily relies on the quality, diversity, and size of the training data. Limited or biased datasets might restrict the model's generalization to new or diverse images.

Linguistic Complexity:

Capturing complex language nuances and context in image descriptions remains a challenge, leading to potential inaccuracies or limited expressiveness in generated captions.

Overfitting and Generalization:

Overfitting to specific image features may hinder the model's ability to generalize to unseen or diverse images, impacting the quality and relevance of generated captions.

Ethical Implications:

Biases inherent in the training data could reflect in generated captions, necessitating

careful handling of ethical concerns related to fairness, inclusivity, and avoiding stereotypes.

Computation and Resource Intensiveness:

The CNN-LSTM architecture can be computationally intensive, requiring significant resources, especially when working with larger datasets or complex models.

Scope and Future Directions:

Incorporating Advanced Architectures:

Exploring hybrid models or advanced architectures beyond CNN-LSTM, such as attention mechanisms, transformers, or multimodal models, could enhance the model's performance.

Ethical AI Integration:

Integrating fairness and bias reduction techniques to ensure ethically sound caption generation, actively addressing biases and ensuring inclusivity in generated content.

5.3.1 Scope of the work

Ethical AI Integration:

Integrating fairness and bias reduction techniques to ensure ethically sound caption generation, actively addressing biases and ensuring inclusivity in generated content.

Multimodal Integration:

Integrating other modalities (such as audio or text) with images could lead to more comprehensive and contextually rich captions, improving overall description quality.

Fine-tuning for Specific Domains:

Tailoring the model for specific domains or applications (e.g., medical imaging, fashion, or art) could enhance its accuracy in generating domain-specific captions.

User-Centric Design and Evaluation:

Implementing user feedback and evaluations to refine the system based on practical use cases, focusing on user preferences and application-specific requirements.

Robustness Enhancement:

Developing techniques to improve the model's robustness to diverse image styles, qualities, and variations, ensuring consistent performance across different scenarios.

Human-in-the-loop Approaches:

Integrating human review or correction systems to refine and enhance the quality of generated captions, ensuring more accurate and contextually relevant outputs.

The project’s scope encompasses the improvement and evolution of the CNN and LSTM-based image caption generator, addressing limitations and exploring advancements for more accurate, contextually rich, and ethically sound captioning systems. The continual exploration of advanced techniques and user-centric enhancements will significantly contribute to the system's development and practical applications.

Chapter 6

Conclusion and Recommendations

6.1 Summary of the Project

The image caption generator is built by using deep learning techniques which are being used to generate an absolute caption for any input images. The use of this technology spans a variety of industries and offers a wide range of features, including automatic graphic presentation, enhanced user engagement, information that to features improved internal flexibility, efficient graphics reference and retrieval, assistive technology for convenience, supported search and analysis, and related

functionality. this deep learning model used many technologies that analyze visual content, including computer vision, natural language processing, convolutional neural networks, and LSTM etc, these technologies are called the concept of descriptive text annotation. The CNN is been used as the most useful and efficient method for caption generators which is being a combination of long-term and short-term memory. By this we can able to generate a clear and coherent caption.

6.2 Contributions and achievements

The image caption generator is built by using deep learning techniques which are being used to generate an absolute caption for any input images. The use of this technology spans a variety of industries and offers a wide range of features, including automatic graphic presentation, enhanced user engagement, information that to features improved internal flexibility, efficient graphics reference and retrieval, assistive technology for convenience, supported search and analysis, and related

functionality. this deep learning model used many technologies that analyze visual content, including computer vision, natural language processing, convolutional neural networks, and LSTM etc, these technologies are called the concept of descriptive text annotation. The CNN is been used as the most useful and efficient method for caption generators which is being a combination of long-term and short-term memory. By this we can able to generate a clear and coherent caption.

6.3 Recommendations for future work

Incorporation of Advanced Architectures:Explore and integrate advanced architectures beyond CNN and LSTM, such as transformer models or attention mechanisms, to further enhance the system's accuracy and context understanding.

Multimodal Integration:Investigate the fusion of multiple modalities (text, audio, etc.) with images to create more comprehensive and contextually rich captions.

Ethical AI Advancements:Continue refining techniques to mitigate biases, ensuring fairness and inclusivity in generated captions, thereby contributing to the ongoing development of responsible AI.

Fine-tuning for Specific Domains:Tailor the model for specific domains or applications, such as medical imaging, fashion, or art, to enhance accuracy and relevance in specialized fields.

User-Centric Design and Evaluation:Implement user feedback mechanisms for iterative improvements, ensuring practical usability and relevance in real-world applications.

Robustness Enhancement:Develop techniques to improve the model's robustness to various image styles, qualities, and variations, ensuring consistent performance across different scenarios.

Human-in-the-loop Approaches:Implement human review or correction systems to refine and enhance the quality of generated captions.

Bibliography

[1] Portet F, Reiter E, Gatt A, Hunter J, Sripada S, Freer Y, Sykes C (2009). "Automatic Generation of Textual Summaries from Neonatal Intensive Care Data" (PDF). Artificial Intelligence.

[2] Ian Goodfellow, Yoshua Bengio, Aaron Courville, Deep Learning, 2016

[3] Richard Szeliski, Computer Vision: Algorithms and Applications, 2010.

[4] Oriol Vinyals, Alexander Toshev, Samy Bengio, D. Computer ScienceIEEE Conference on Computer Vision and Pattern…2015

[5] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4651–4659, 2016.

[6] R. Subash (November 2019): Automatic Image Captioning Using Convolution Neural Networks and LSTM.

[7] Simao Herdade, Armin Kappeler, Kofi Boakye, Joao Soares (June 2019): Image Captioning: Transforming Objects into words.

[8] Long Short-Term Memory, Hochreiter, Sepp and Schmidhuber, Jurgen, 1997.

[9] Gers, F. A.; Schmidhuber, J. & Cummins, F. A. (2000), 'Learning to Forget: Continual Prediction with LSTM.', NeuralComputation 12 (10) , 2451-2471 .

[10]Venkatesan, Ragav; Li, Baoxin (2017-10-23). Convolutional Neural Networks in Visual Computing: A Concise Guide. CRC Press. ISBN 978-1-351-65032-8.

Appendices

Appendix A

Source code

The code also includes training steps, which involve teaching the model to improve its captioning abilities by exposing it to many images and captions, allowing it to learn from its mistakes. Once trained, the code can take a new image and use the model to generate descriptive captions. Additionally, there are components for handling words and converting them between text and numerical representations.

To assess the quality of generated captions, the code may include checks using metrics such as BLEU, METEOR, CIDEr, and ROUGE. It also provides settings and options that allow users to tweak parameters like learning speed, batch size, and model complexity. Finally, there might be code for saving the model's learned knowledge and loading it later to avoid retraining from scratch. Depending on the specific application, there could be extra features, such as user interfaces or web app integrations to make the image captioning experience more user-friendly.

CODE:

```
import  string

import numpy as np

from PIL import Image

import os
```

```
from pickle import dump, load
import numpy as np
import tensorflow as tf
from tensorflow.keras.utils import to_categorical
from tensorflow.keras.layers import add
from tensorflow.keras.models import Model, load_model
from tensorflow.keras.layers import Input, Dense, LSTM, Embedding, Dropout
from keras.applications.xception import Xception, preprocess_input
from keras.preprocessing.image import load_img, img_to_array
from keras.preprocessing.text import Tokenizer
from keras.preprocessing.sequence import pad_sequences
# small library for seeing the progress of loops.
from tqdm.notebook import tqdm
tqdm.pandas()
from tensorflow.keras.utils import plot_model


# define the captioning model
def define_model(vocab_size, max_length):

    # features from the CNN model squeezed from 2048 to 256 nodes
    inputs1 = Input(shape=(2048,))
    fe1 = Dropout(0.5)(inputs1)
```

```
    fe2 = Dense(256, activation='relu')(fe1)

    # LSTM sequence model
    inputs2 = Input(shape=(max_length,))
    se1 = Embedding(vocab_size, 256, mask_zero=True)(inputs2)
    se2 = Dropout(0.5)(se1)
    se3 = LSTM(256)(se2)

    # Merging both models
    decoder1 = add([fe2, se3])
    decoder2 = Dense(256, activation='relu')(decoder1)
    outputs = Dense(vocab_size, activation='softmax')(decoder2)

    # tie it together [image, seq] [word]
    model = Model(inputs=[inputs1, inputs2], outputs=outputs)
    model.compile(loss='categorical_crossentropy', optimizer='adam')
    # summarize model
    print(model.summary())
    plot_model(model,to_file='/content/drive/MyDrive/ML/model.png',
show_shapes=True)

    return model
```

Appendix B

Screen shots

B.1 Output Of The Project

To generate output data for an image caption generator using a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) model, you would typically need to provide an image as input, and the model would generate a textual caption for that image. The actual output data would depend on the specific dataset and images you are working with, as well as the design and training of your CNN-LSTM model. The caption can vary in length and complexity, and the model's performance would also affect the quality of the generated captions. You would typically have a dataset of images paired with their corresponding captions to train and evaluate the model.

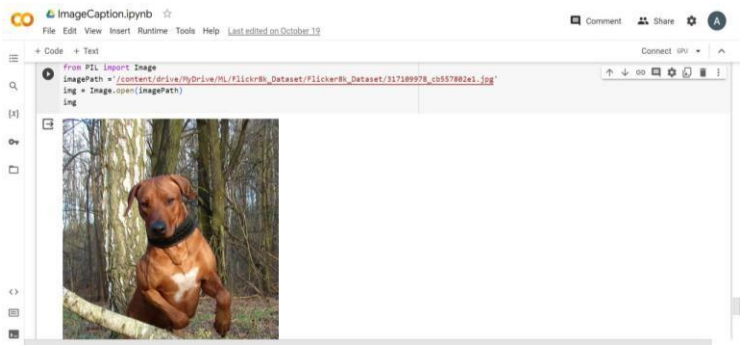


Figure B.1: output

Appendix C
Data sets used in the Project

There are several datasets, we'll be combining them in this project. Flickr_8k_text and Flickr8k are two datasets that have the same main purpose. The programming language will be using Python language. Python is used because of its vast libraries and it will make it easy for developers to understand and write code.

Data visualization is used in the image caption generator. we required different components like NLP, Deep Learning, RNN, CNN, LSTM, etc in image caption generators. The certain requirements Collected into the Flickr8k dataset are 8000 photos. For data analysis, we gathered information from a variety of Flickr groups and mixed up the order of the words. Python language is capable of visualization tasks. Jupyter Notebooks are our chosen tool

The dataset has a number of characteristics of the main goals of our project:

In this dataset, the model is created by including a different number of labels in a single image which helps in overfitting. The dataset increases the variety of image categories and increases the image annotation model's adaptability. The model's strength is increased by the variety of image categories used for training. The flicker8k_Dataset and Flickr_8k_text are available datasets in websites and in that dataset, it has already been filled with data.

8091 photos can be found in the dataset folder called flicker8k_Dataset.

Flickr_8k_text is a dataset folder that houses text documents and image captions.