

# Multilingual Image Captioning with Voice Assistance

*A Project Report submitted by*  
**Krishna Kumari Ravuri**

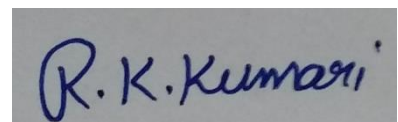
*in partial fulfilment of the requirements for the award of the degree of*  
**M.Tech.**



**Indian Institute of Technology Jodhpur**  
**Department of Mathematics and**  
**Computational Science**  
*November 2023*

## Declaration

I hereby declare that the work presented in this Project Report titled Multilingual Image Captioning with Voice Assistance submitted to the Indian Institute of Technology Jodhpur in partial fulfilment of the requirements for the award of the degree of M.Tech., is a bonafide record of the research work carried out under the supervision of Professor Dr. Sukhendu Ghosh. The contents of this Project Report in full or in parts, have not been submitted to and will not be submitted by me to any other Institute or University in India or abroad for the award of any degree or diploma.

A rectangular box containing a handwritten signature in blue ink that reads "R. K. Kumari".

Signature  
*KRISHNA KUMARI RAVURI*  
M22AI567

## Certificate

This is to certify that the Project Report titled Multilingual Image Captioning with Voice Assistance, submitted by Krishna Kumari Ravuri (M22AI567) to the Indian Institute of Technology Jodhpur for the award of the degree of M.Tech., is a bonafide record of the research work done by her under my supervision. To the best of my knowledge, the contents of this report, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Signature  
Dr. Sukhendu Ghosh  
Associate Professor  
Department of Mathematics  
IIT Jodhpur, Rajasthan, India

## **Acknowledgment**

I am incredibly thankful to my supervisor Professor Dr. Sukhendu Ghosh, Department of Mathematics, Indian Institute of Technology, Jodhpur, for his motivation and tireless efforts to help me gain profound knowledge of the research area and for supporting me throughout the life cycle of my M.Tech. dissertation work. Significantly, the extensive comments, healthy discussions, and fruitful interactions with the supervisor directly impacted the final form and quality of M. Tech. dissertation work.

My deepest regards to my Parents for their blessings, affection, and continuous support. Also, Last but not least, I thank GOD, the almighty, for giving me the inner willingness, strength, and wisdom to carry out this research work successfully

Signature  
Dr. Sukhendu Ghosh  
Associate Professor  
Department of Mathematics  
IIT Jodhpur, Rajasthan, India

## **Abstract**

This project endeavour to address the significant educational and employment challenges faced by the visually impaired population in India due to limited access to Braille and expensive assistive technologies. With less than 10 percent of blind individuals having access to Braille education, a critical literacy gap exists. To mitigate this, our proposal outlines the development of an affordable and accessible solution utilizing computer vision, object recognition, and speech synthesis technologies. The envisioned model aims to convert text from images into machine-readable format and subsequently employ speech synthesis to relay real-time information about the user's surroundings. A key focus is the incorporation of regional languages, given India's linguistic diversity, ensuring the model's capability to read aloud text in multiple languages. This comprehensive approach seeks to enhance the quality of life and open up new opportunities for the visually impaired community in India.

**Keywords:** Computer vision, Object recognition, Speech synthesis, Multilingual text recognition, Machine learning , convolutional neural network (CNN), recurrent neural network (RNN)

<b>Content</b>	<b>Page</b>
Abstract	1
Introduction	3
Background	4
Literature Survey	5
Problem Definition	7
Objective	8
Technology Used	9
Dataset Specification	11
Data Pre-processing	14
Methodology	24
Results	36
Future Scope	49
References	50

## 1. INTRODUCTION

In the vast tapestry of India's societal landscape, the visually impaired encounter substantial challenges in the realms of education and employment, primarily stemming from a glaring lack of accessible resources. The deficiency in Braille education, coupled with the exorbitant costs associated with assistive technologies, has resulted in a poignant literacy gap, leaving less than 10 percent of blind individuals with the means to learn Braille.

One distinctive aspect of our project lies in its commitment to linguistic diversity. India, with its multitude of languages and dialects, presents a unique challenge and opportunity. Recognizing this, our model is meticulously designed to accommodate and proficiently handle multiple regional languages. Beyond the technological intricacies, our emphasis on linguistic inclusivity is a conscious effort to ensure that the model can seamlessly convert and articulate text in diverse languages, resonating with the linguistic preferences of the user.

This linguistic adaptability is not merely a feature but a foundational principle that aims to foster inclusivity and usability across India's diverse linguistic spectrum. This adaptability resonates with the linguistic preferences of users, transcending technological intricacies to promote inclusivity across India's expansive linguistic spectrum.

At the core of our solution resides a sophisticated technological framework seamlessly integrating multiple components. Employing a pre-trained convolutional neural network (CNN), specifically InceptionV3, empowers our model with robust image feature extraction. Complementing this is an attention mechanism that enhances the model's focus on pertinent aspects of the input image during caption generation. The recurrent neural network (RNN)-based decoder further refines the process, generating captions based on the extracted features to construct a coherent and contextually relevant description of the user's environment.

This holistic technological approach not only addresses immediate challenges faced by the visually impaired but positions our solution as a scalable and adaptive assistive tool for the future. Through this fusion of technological prowess and inclusivity, our project endeavours to propel meaningful advancements in the educational and employment landscape for the visually impaired community in India.

## **2. BACKGROUND**

impaired population in India, particularly in the realms of education and employment. Existing barriers, such as limited access to Braille resources and the high costs associated with assistive technologies, contribute to a pronounced literacy gap among the visually impaired, with fewer than 10 percent having the opportunity to learn Braille. Recognizing the critical need for innovative solutions, this project seeks to leverage advancements in computer vision, object recognition, and speech synthesis technologies to create a transformative tool.

The motivation to address these challenges is rooted in a broader commitment to inclusivity and empowerment. The dearth of accessible resources and the lack of affordable assistive technologies not only limit educational opportunities but also impede access to meaningful employment for the visually impaired. This project aims to fill this void by developing a sophisticated yet accessible model that harnesses the power of technology to enhance the quality of life for the visually impaired community in India.

Furthermore, the linguistic diversity of India adds a layer of complexity to this challenge. The project acknowledges and seeks to address the rich tapestry of languages and dialects spoken across the country. By designing the model to read aloud text in multiple regional languages, the project aims to ensure that the solution is not only technologically robust but also culturally and linguistically sensitive.

In the broader context of technological advancements and social responsibility, this project envisions contributing to the creation of a more inclusive and equitable society. By bridging the educational and employment gaps for the visually impaired in India, the project aligns with a commitment to leveraging technology for social good and empowering marginalized communities. Through a combination of cutting-edge technology and a deep understanding of societal challenges, the project endeavours to make a meaningful impact on the lives of the visually impaired in India.



### 3. LITERATURE SURVEY

#### 1. Global Overview:

The World Health Organization's "World Report on Vision" provides a comprehensive global perspective on visual impairment, offering insights into the challenges faced by visually impaired individuals worldwide. This report is instrumental in understanding the broader context of visual impairment and initiatives on a global scale.

#### 2. Assistive Technologies Worldwide:

The RNIB's "Technology for Life: Assistive Technology Report" is a valuable resource for exploring successful global models of assistive technologies. It highlights innovations and best practices that have proven effective in enhancing the lives of visually impaired individuals, serving as a benchmark for the proposed project.

#### 3. Indian Perspective on Assistive Technologies:

The academic paper by George and Mathew (2018) delves into "Assistive Technologies in Education for Visually Impaired: An Indian Perspective." This work specifically addresses the Indian context, providing insights into existing technologies, challenges faced, and potential areas for improvement in education.

#### 4. Saksham's Technology Initiatives:

"Saksham: Technology for the Blind" offers on-the-ground insights into technology initiatives for the blind in India. The organization's website may provide real-world case studies and user testimonials, contributing to a better understanding of the local landscape and potential project applications.

#### 5. Multilingual Text Recognition Systems:

Singh and Yadav's (2020) academic review on "Multilingual Text Recognition Systems" explores the landscape of systems crucial for the project's goal of regional language adaptability. It provides insights into challenges and solutions in recognizing and vocalizing text in diverse languages.

#### 6. ICT Accessibility in Indian Languages:

The resource from the Centre for Internet and Society (2018) titled "Accessibility of ICTs in Indian Languages" offers an in-depth exploration of the accessibility of Information and Communication Technologies in Indian languages. It sheds light on challenges related to language diversity and the development of inclusive technologies.

#### 7. Ethical Considerations in Assistive Technologies:

Foley and Mathews (2019) discuss "Ethical Considerations in the Use of Assistive Technology for People with Disabilities." This academic paper guides the project in ensuring ethical principles, particularly regarding privacy and dignity, are upheld in the development of the proposed solution.

#### 8. Be My Eyes Case Study:

"Be My Eyes," a real-world implementation providing remote assistance to visually impaired individuals, serves as a relevant case study. This application showcases the potential of technology to connect and improve the lives of visually impaired individuals.

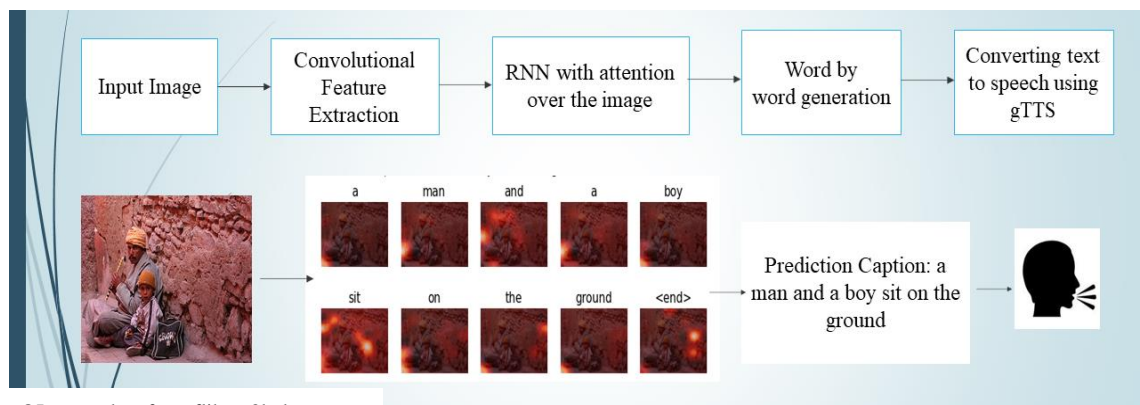
## 9. Language Processing for Indian Languages:

Ghosal and Hazarika's (2019) academic review on "Language Processing for Indian Languages" provides insights into the challenges and advancements in linguistic diversity. It informs the project's approach to handling multiple regional languages.

## 10. Microsoft Translator for Education:

The Microsoft Translator for Education offers a practical tool for language adaptability. This resource aligns with the project's goal of enhancing educational opportunities for the visually impaired through translation services in multiple languages.

This condensed literature survey synthesizes a range of global and local perspectives, technology initiatives, case studies, and academic reviews that collectively inform the proposed project on developing assistive technology for the visually impaired in India.



## **4. PROBLEM DEFINITION**

The visually impaired population in India encounters formidable challenges in accessing education and employment opportunities due to a critical lack of accessible technologies. Less than 10 percent of blind individuals in India have the opportunity to learn Braille, resulting in a substantial literacy gap. This stark reality limits their access to information and inhibits their ability to navigate their surroundings independently. The absence of affordable and inclusive assistive technologies exacerbates the difficulties faced by the visually impaired, hindering their quality of life and impeding their integration into educational and professional spheres. To address these pressing issues, there is a crucial need for an innovative solution that combines computer vision, object recognition, and speech synthesis technologies. This project aims to develop a model capable of converting visual information into audible text, providing real-time environmental insights and thereby enhancing the overall quality of life and opportunities for the visually impaired in India. The key challenge lies in ensuring the model's adaptability to India's linguistic diversity, making it proficient in converting images to text and vocalizing content in multiple regional languages. By tackling these challenges, the project endeavours to bridge the existing gaps, empowering the visually impaired community to navigate their surroundings more effectively and participate more fully in educational and professional endeavours.

## 5. OBJECTIVE

Design and implement an affordable and accessible assistive technology solution leveraging computer vision, object recognition, and speech synthesis technologies. The primary objective is to create a tool that overcomes the financial barriers faced by the visually impaired in accessing advanced assistive technologies.

Enable the model to process visual information in real-time, providing the visually impaired with immediate insights into their surroundings. This includes reading text aloud, identifying objects, and offering contextual information to enhance their overall environmental awareness.

Ensure the model's proficiency in converting images to text and reading out information in multiple regional languages. The objective is to make the technology inclusive and culturally relevant for the diverse linguistic landscape of India, enhancing its usability and impact.

Facilitate better access to education for the visually impaired by providing real-time audio feedback on written content. The model should be capable of reading textbooks, study materials, and other educational resources aloud, thereby promoting inclusive learning environments.

Support the visually impaired in their pursuit of employment by providing real-time information about their work environment. This includes reading written documents, identifying objects, and offering audio cues to enhance their workplace navigation and productivity.

Prioritize a user-centric design, focusing on human-computer interaction principles to create an intuitive and user-friendly experience. The objective is to empower visually impaired users with a technology solution that is easy to navigate and seamlessly integrates into their daily lives.



## **6. TECHNOLOGY USED**

### **Computer Vision**

InceptionV3 (Convolutional Neural Network): InceptionV3, a pre-trained deep learning model, is employed for image feature extraction. The model analyses visual input, extracting high-level features from images to comprehend and interpret their content.

### **Object Recognition**

Object recognition is integrated to identify and recognize objects within the visual input, enhancing the contextual understanding of the user's environment.

### **Speech Synthesis**

Speech synthesis technology converts generated textual captions into audible information for real-time feedback to visually impaired users.

### **Machine Learning**

Convolutional Neural Network (CNN):

CNNs are used for image processing and feature extraction. The pre-trained InceptionV3 model, a specific instance of a CNN, effectively understands complex visual features.

Recurrent Neural Network (RNN):

RNNs are employed for natural language processing. The RNN-based decoder refines caption generation, producing coherent and contextually relevant descriptions based on extracted image features.

### **Attention Mechanism**

An attention mechanism is incorporated to enhance focus on pertinent aspects of the input image during caption generation, contributing to more accurate and contextually rich descriptions.

### **Google Translator**

Google Translator is used for the translation of generated captions into multiple regional languages, ensuring linguistic inclusivity.

### **Beam Search**

Beam search is employed as a post-processing step to refine and select the most probable captions, enhancing overall accuracy and coherence.

## **Jupyter Notebook**

Jupyter Notebook serve as the development environment for running the caption generator, providing an interactive interface for coding, visualization, and documentation.

## **Python**

Python is the chosen programming language for its versatility and extensive libraries, suitable for developing complex systems involving machine learning and deep learning.

These technologies collectively form a sophisticated and adaptive assistive technology solution, addressing challenges in education, employment, and linguistic diversity for the visually impaired community in India. The integration of computer vision, natural language processing, and translation services reflects a comprehensive approach to improving the quality of life for the target users.

## **JarvisLab**

JarvisLab is an advanced research environment specializing in artificial intelligence (AI) and machine learning (ML) development. It offers a suite of tools, libraries, and computing resources tailored for AI projects, including deep learning models, natural language processing (NLP) algorithms, computer vision applications, and more. JarvisLab is designed to facilitate research, experimentation, and deployment of cutting-edge AI solutions, making it a preferred choice for researchers, engineers, and data scientists working on complex AI projects.

## 7. DATASET SPECIFICATIONS

### Flickr\_8K Dataset

The Flickr\_8K dataset is a widely used resource for training and evaluating image caption generators in the field of computer vision and natural language processing.

#### Dataset Overview:

Number of Images: 8,091 out of which used 2000 images used for partial presentation.

Image Content: Diverse scenes, objects, and activities.

#### File Structure:

Images Folder: Contains JPEG image files with unique identifiers.

Captions text file: Holds text files with captions for each image.

#### File Naming:

Images in 'Flicker8k\_Dataset' are named using unique identifiers.

Captions are associated with images in 'captions.txt'.

#### Caption Information:

Token File: 'captions.txt' includes image names paired with their captions.

Captions per Image: Typically, five captions per image.



A child in a pink dress is climbing up a set of stairs in an entry way  
A girl going into a wooden building  
A little girl climbing into a wooden playhouse  
A little girl climbing the stairs to her playhouse  
A little girl in a pink dress going into a wooden cabin

©Images taken from flicker\_8k dataset



A little girl covered in paint sits in front of a painted rainbow with her hands in a bowl  
A little girl is sitting in front of a large painted rainbow  
A small girl in the grass plays with fingerpaints in front of a white canvas with a rainbow on it  
There is a girl with pigtails sitting in front of a rainbow painting  
Young girl with pigtails painting outside in the grass

The Flickr8k dataset is a comprehensive resource for training and evaluating models in the domains of computer vision and natural language processing.

### **Image Collection:**

The dataset consists of a diverse collection of 8,000 high-resolution images sourced from Flickr, a popular photo-sharing platform. These images cover a wide range of scenes, objects, and contexts, providing a rich variety of visual content for analysis and modelling.

### **Annotations:**

Each image in the dataset is associated with a set of five textual descriptions or captions, resulting in a total of 40,000 captions across the entire dataset. These captions are manually created by human annotators and aim to describe the content of the corresponding image in natural language.

### **Multimodal Data:**

The Flickr8k dataset is multimodal, meaning it combines visual data (images) with textual data (captions). This makes it particularly suitable for tasks that involve understanding and generating descriptions based on both visual and textual inputs, such as image captioning, visual question answering (VQA), and multimodal retrieval.

### **Training and Testing Split:**

The dataset is commonly divided into training and testing sets to facilitate model training, validation, and evaluation. The standard split involves 6,000 images for training and 2,000 images for testing/validation. This separation ensures that models are trained on one subset of data and tested on another, providing a fair assessment of their generalization ability.

### **File Format:**

Typically, the Flickr8k dataset is distributed in a structured format. This includes a directory containing the image files (e.g., in JPEG format) and accompanying text files that list the image IDs along with their respective captions. This format enables easy access and processing of both visual and textual data.

## **COCO DATASET:**

The COCO (Common Objects in Context) dataset is a widely used benchmark dataset in computer vision, particularly for tasks like object detection, segmentation, and captioning. Here is a detailed overview of the COCO dataset:

### **Image Collection:**

The COCO dataset comprises a large collection of images covering a wide range of scenes, objects, and contexts. It contains over 200,000 images in total, making it one of the largest datasets for object recognition and scene understanding tasks.



### **Annotations:**

Each image in the COCO dataset is densely annotated with multiple types of annotations, including:

**Object Instances:** Annotations outlining the boundaries of individual objects within the image.

**Object Categories:** Labels specifying the category or class of each annotated object (e.g., person, car, dog, etc.).

**Keypoints:** Annotations marking specific keypoints on objects (e.g., keypoints on human bodies).

**Captions:** Descriptive textual captions that describe the content of the image in natural language.

### **Multimodal Data:**

The COCO dataset is multimodal, incorporating both visual data (images) and textual data (captions). This multimodal nature enables a wide range of research and applications, including image captioning, object detection, semantic segmentation, and human pose estimation.

### **Task-specific Subsets:**

Within the COCO dataset, there are task-specific subsets, such as COCO-Stuff (for semantic segmentation), COCO Captioning (for image captioning), and COCO Keypoints (for human pose estimation). These subsets provide focused data for particular research areas while maintaining consistency with the overall COCO dataset.

### **File Format:**

The dataset is typically provided in a structured format, including image files (e.g., JPEG) along with annotation files in formats like JSON. These annotation files contain detailed information about object annotations, categories, keypoints, and captions associated with each image.

The COCO dataset's richness in annotations, diverse visual content, and multimodal nature make it a valuable resource for advancing research in computer vision, machine learning, and related fields. Its widespread adoption has contributed significantly to the development of state-of-the-art models and techniques across various visual understanding tasks.

## 8. DATA PRE-PROCESSING

### 8.1 Data collection

#### 8.1.1 Collecting the Images by Class

In my project using the COCO (Common Objects in Context) dataset, I've opted for a focused approach by selecting a subset of 16 classes from the dataset for my specific tasks and analyses. This decision was based on the objectives and requirements of my project, ensuring that the chosen classes align closely with the intended goals.

By narrowing down the classes to this targeted set of 16, I aim to streamline my efforts, reduce computational complexity, and tailor my algorithms or models to effectively handle object recognition or segmentation tasks related to these specific classes. This focused strategy allows me to gain deeper insights into the characteristics and nuances of the selected object categories, enabling more accurate and meaningful analyses within the COCO dataset framework.

Overall, my choice to concentrate on these 16 classes within the COCO dataset reflects a deliberate and strategic approach aimed at optimizing resource utilization, improving model performance, and deriving valuable insights relevant to my research or application domain. Implemented Python code to collect images of specific classes from the COCO dataset based on their captions.

#### **Importing Libraries:**

Imported necessary libraries such as `os` for file operations, `shutil` for copying files, and `pycocotools.coco` for working with the COCO dataset.

#### **Setting up Paths:**

It defines the path to the COCO dataset folder (`data_dir`) and the annotation file (`annotation_file`) containing captions. It also specifies the directory containing the images (`image_dir`).

#### **Initializing COCO API:**

The COCO API is initialized using the annotation file to work with captions and image metadata.

#### **Selecting Images:**

It selects a range of image IDs from the dataset (specified by `start_index` and `end_index`) to process. In your code, it selects the next 100,000 images starting from index 1.

#### **Creating Folders based on the class:**

It creates a folder (`selected_folder`) to store selected images and a subfolder (`class_folder`) for images belonging to a specific class (e.g., "dog" in your code).

#### **Copying Images:**

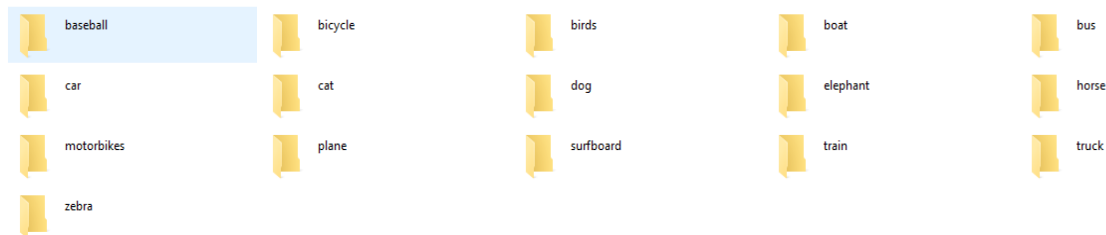
For each selected image ID, the code retrieves the image's file name and path. It then checks if any of the image's captions contain the specified class name (e.g., "dog"). If a caption contains the class name, the image is copied to the corresponding class folder (`class_folder`), and its captions are saved to a text file inside that folder.

## Caption Filtering:

The code uses a conditional check (if any('dogs ' in caption.lower() for caption in captions)) to filter images based on captions containing the specified class name. You can modify this check to target different classes or criteria as needed.

## Saving Captions:

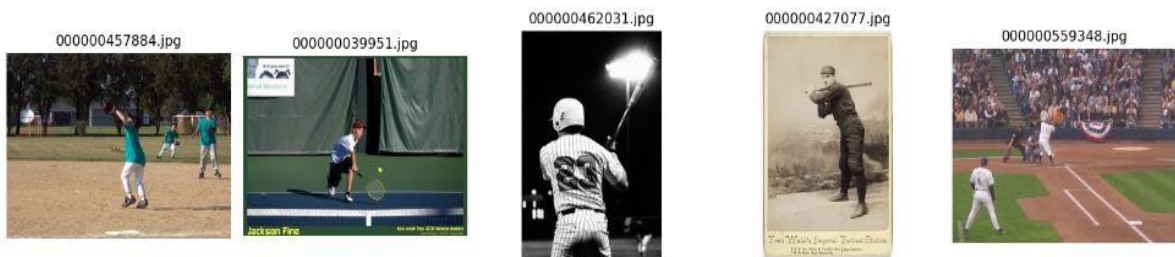
Captions associated with selected images are saved to text files with the format <image\_id>\_captions.txt inside the respective class folders.



- Class 'baseball': 2,441 images
- Class 'bicycle': 973 images
- Class 'birds': 1,576 images
- Class 'boat': 1,561 images
- Class 'bus': 2,127 images
- Class 'car': 3,155 images
- Class 'cat': 2,794 images
- Class 'dog': 3,537 images
- Class 'elephant': 1,126 images
- Class 'horse': 1,518 images
- Class 'motorbikes': 150 images
- Class 'plane': 2,255 images
- Class 'surfboard': 1,720 images
- Class 'train': 2,651 images
- Class 'truck': 1,404 images
- Class 'zebra': 1,405 images
- The total number of images across all classes is 30,393.

## Random Images Display Class wise:

Displaying random images from class 'baseball':



Displaying random images from class 'bicycle':



Displaying random images from class 'birds':



Displaying random images from class 'boat':



Displaying random images from class 'bus':



Displaying random images from class 'car':



Displaying random images from class 'cat':



Displaying random images from class 'dog':





Displaying random images from class 'elephant':



Displaying random images from class 'horse':



Displaying random images from class 'motorbikes':



Displaying random images from class 'plane':

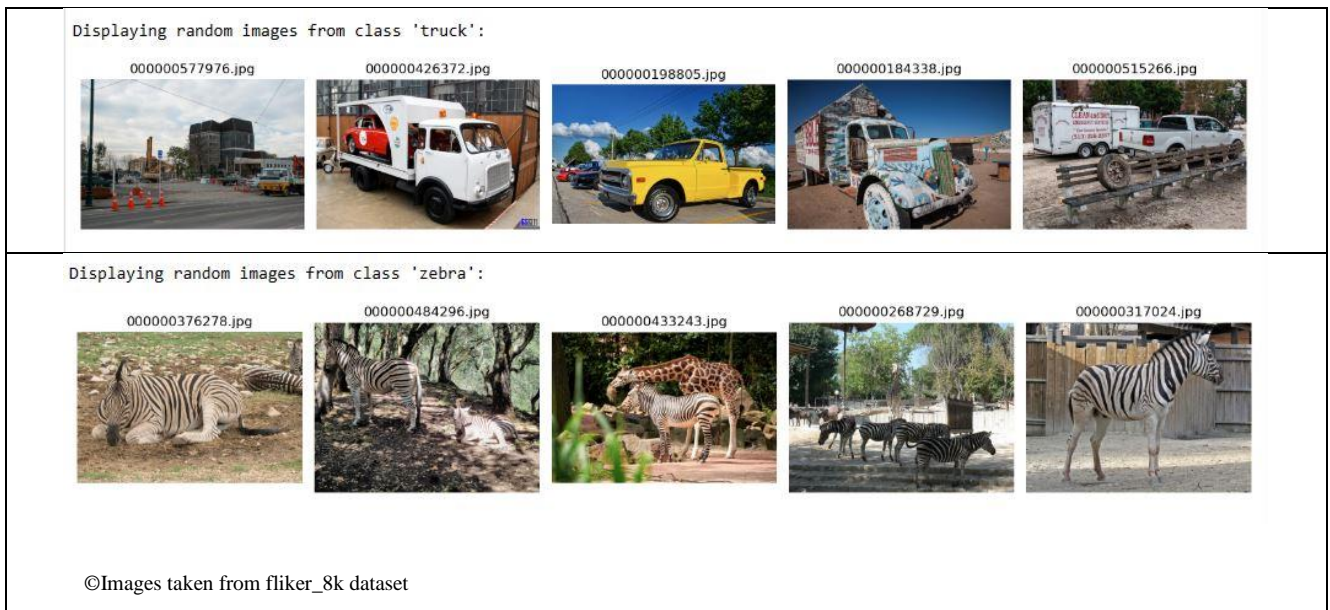


Displaying random images from class 'surfboard':



Displaying random images from class 'train':





### 8.1.1. Merging Image Data and Corresponding Captions:

Merging image data and corresponding captions from multiple folders into a cohesive dataset. It begins by defining the directory paths for input (containing image folders) and output (for the merged dataset). Subsequently, it creates necessary subfolders within the output directory, such as one specifically for images.

The core functionality revolves around iterating through each folder in the input directory. For every folder encountered, the script counts the number of images and proceeds to process each image file. If an image has a .jpg extension, it is copied to the designated output image folder. Simultaneously, the script retrieves captions for each image by referencing corresponding caption files named using image IDs followed by '\_captions.txt'. Captions are then extracted using a function (read\_captions) and stored in a dictionary, associating image filenames with their respective caption lists. To maintain consistency and clarity, captions exceeding 5 lines are trimmed during processing.

Upon completing the image and caption extraction process for all folders, the script compiles this information into a single text file within the output folder. This file lists each image filename alongside its corresponding caption, adhering to a structured format conducive to subsequent analysis or model training.

## 8.2 Pre-Processing the Captions:

### 8.2.1. Create a data frame which summarizes the image, path & captions as a data frame

Read image captions from a file, extracts image IDs, constructs image paths, and creates a DataFrame summarizing the image IDs, paths, and captions. This structured DataFrame is useful for organizing and analysing image data along with their associated textual descriptions in a tabular format.

	ID	Path	Captions
0	1000268201_693b08cb0e.jpg	/datasets/flickr/Images/1000268201_693b08cb0e.jpg	A child in a pink dress is climbing up a set of stairs in an entry way
1	1000268201_693b08cb0e.jpg	/datasets/flickr/Images/1000268201_693b08cb0e.jpg	A girl going into a wooden building
2	1000268201_693b08cb0e.jpg	/datasets/flickr/Images/1000268201_693b08cb0e.jpg	A little girl climbing into a wooden playhouse
3	1000268201_693b08cb0e.jpg	/datasets/flickr/Images/1000268201_693b08cb0e.jpg	A little girl climbing the stairs to her playhouse
4	1000268201_693b08cb0e.jpg	/datasets/flickr/Images/1000268201_693b08cb0e.jpg	A little girl in a pink dress going into a wooden cabin
...	...	...	...
40450	997722733_0cb5439472.jpg	/datasets/flickr/Images/997722733_0cb5439472.jpg	A man in a pink shirt climbs a rock face
40451	997722733_0cb5439472.jpg	/datasets/flickr/Images/997722733_0cb5439472.jpg	A man is rock climbing high in the air
40452	997722733_0cb5439472.jpg	/datasets/flickr/Images/997722733_0cb5439472.jpg	A person in a red shirt climbing up a rock face covered in assist handles
40453	997722733_0cb5439472.jpg	/datasets/flickr/Images/997722733_0cb5439472.jpg	A rock climber in a red shirt
40454	997722733_0cb5439472.jpg	/datasets/flickr/Images/997722733_0cb5439472.jpg	A rock climber practices on a rock climbing wall

40455 rows x 3 columns

### 8.2.2. Adding the <start> & <end> token to all those captions

Adding <start> and <end> tokens to captions is a common practice in natural language processing tasks, especially in tasks like image captioning.

**Sequential Modeling:** Neural networks, particularly sequence models like recurrent neural networks (RNNs) or transformers, often require an explicit indication of the start and end of a sequence. This helps the model understand the sequence boundaries and learn the structure of the data more effectively.

**Training Stability:** Including <start> and <end> tokens makes the training process more stable and reliable. It provides clear markers for the model to start generating captions and signals the end of the caption generation process.

**Consistency:** By adding these tokens, you ensure consistency in how captions are represented across different models or experiments. It standardizes the format and makes it easier to preprocess the data uniformly

```
[ '<start> Two men playing baseball in a field on a sunny day <end>',
  '<start> two baseball players are playing baseball on a field <end>',
  '<start> A couple of men play baseball and the batter runs for base <end>',
  '<start> Two guys playing baseball <end>',
  '<start> Batter preparing to swing at pitch during major game <end>']
```

### 8.2.3 Create the vocabulary & the counter for the captions

creating a vocabulary and counting word occurrences in captions are foundational steps in natural language processing and text analysis. They provide essential insights into the linguistic characteristics of the dataset, support data preprocessing, and inform the design and training of models that work with textual data, contributing to the overall effectiveness and accuracy of text-based tasks.

## Vocabulary Creation:

**Understanding Word Diversity:** By creating a vocabulary, you can determine the diversity of words present in the captions. This helps in understanding the range of language used and the richness of textual descriptions associated with the images.

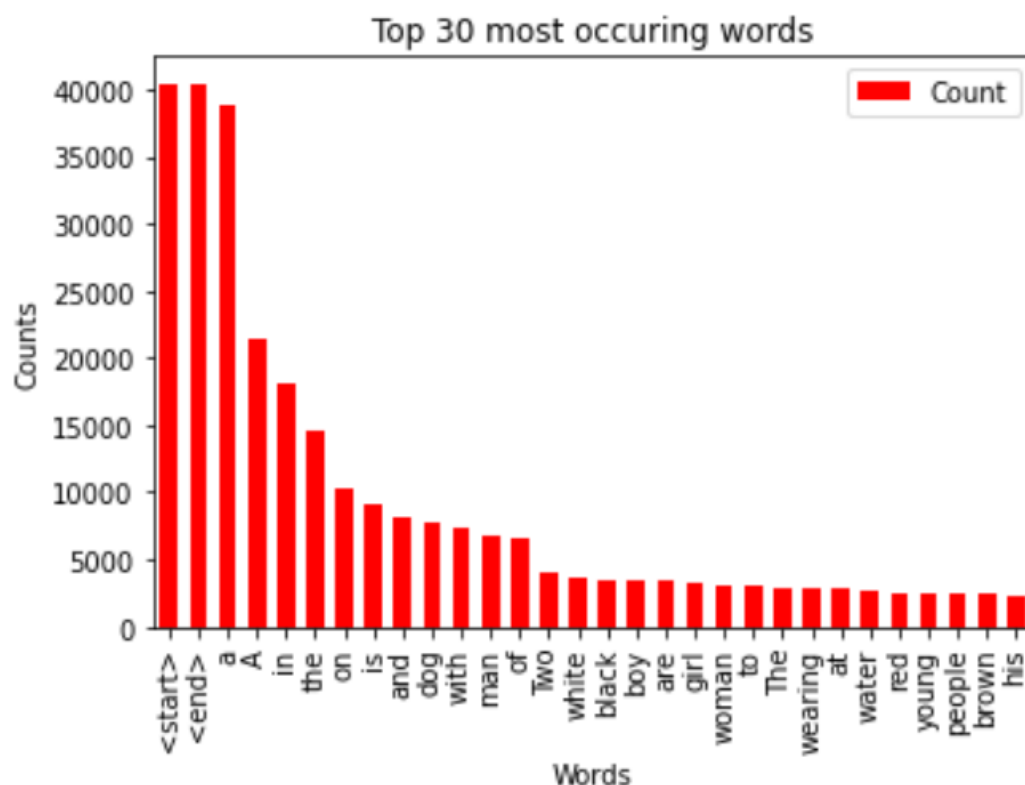
**Identifying Unique Words:** The vocabulary allows you to identify unique words that contribute to the descriptive content of captions. This information is crucial for text analysis tasks and can guide preprocessing steps like filtering out rare or irrelevant words.

## Word Counting:

**Frequency Analysis:** Counting the occurrences of each word provides a frequency analysis, highlighting which words are used most frequently in the dataset. This information is essential for tasks like keyword extraction, identifying common themes, and understanding the predominant topics or objects depicted in the images.

**Data Preprocessing:** Word counts help in data preprocessing by identifying stopwords (commonly used words with little semantic value) and rare words that may need special handling during text processing tasks such as normalization or tokenization.

**Model Training:** In tasks like image captioning where captions are generated by models, word counts can inform the design of the vocabulary used by the model. They guide decisions on vocabulary size, handling of out-of-vocabulary words, and strategies for dealing with word frequencies during training.





## 8.2.4 Creating the tokenized vectors by tokenizing the captions

### Creating Tokenized Vectors:

**Tokenization:** Tokenizing the captions involves splitting them into individual words or tokens. This step is essential for processing textual data and creating a vocabulary of unique words in the dataset.

**Filtering:** Other filters may be applied during tokenization, such as removing punctuation marks, converting words to lowercase, and handling special characters or numbers. These filters help standardize the text and improve the quality of the vocabulary.

**Limiting Vocabulary:** Keeping the total vocabulary to the top 5,000 words helps save memory and computational resources. It focuses on the most relevant and frequently occurring words in the dataset, which are more likely to contribute significantly to the model's understanding and performance.

### Replacing Other Words with "UNK" Token:

**Unknown Token:** Replacing less frequent or out-of-vocabulary words with the "UNK" token (short for "unknown") helps simplify the vocabulary and reduce its size further. This strategy ensures that the model encounters familiar words during training, enhancing its ability to generalize and handle unseen words during inference.

### Creating Word-to-Index and Index-to-Word Mappings:

**Word-to-Index Mapping:** Assigning a unique index to each word in the vocabulary creates a mapping between words and their corresponding numerical representations. This mapping is crucial for encoding textual data into numerical sequences that machine learning models can process.

**Index-to-Word Mapping:** Similarly, creating an index-to-word mapping allows for decoding numerical sequences back into human-readable text. This mapping is essential for interpreting model predictions and generating meaningful outputs.

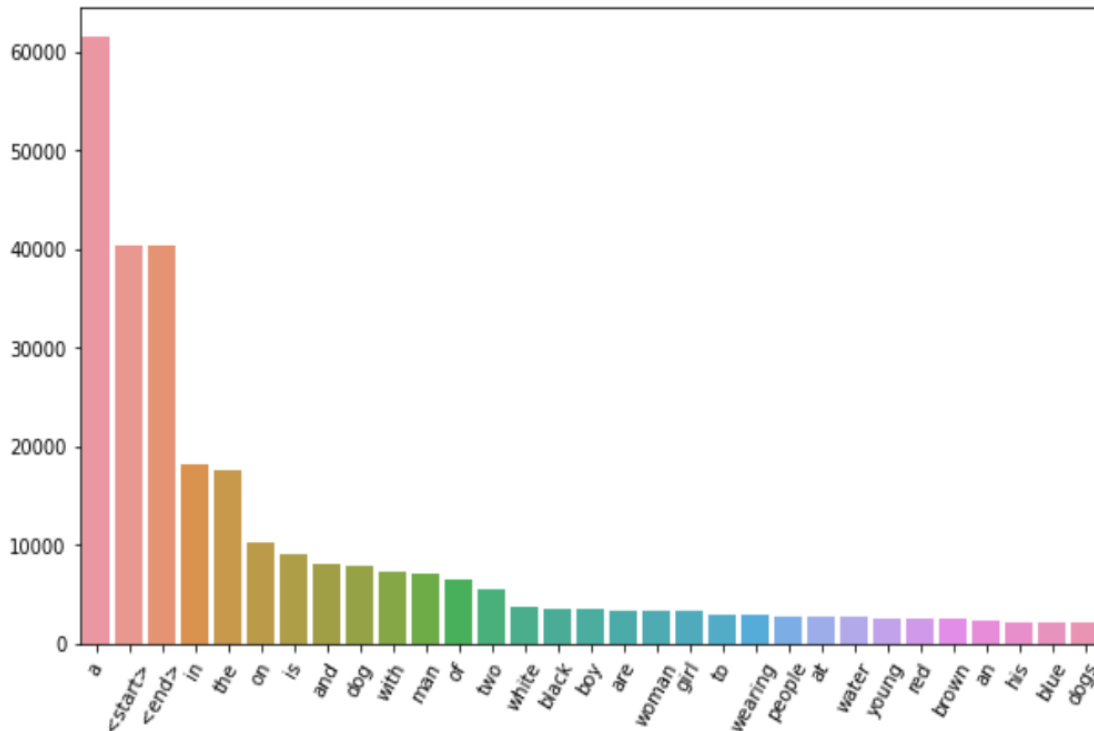
### Padding Sequences:

**Uniform Length:** Padding all sequences to be the same length as the longest one ensures that the input data is uniform and compatible with models that require fixed-length input sequences, such as recurrent neural networks (RNNs) or transformers.

**Batch Processing:** Uniform length sequences facilitate efficient batch processing during model training, where batches of data with consistent dimensions can be processed in parallel, improving computational performance.

**Avoiding Information Loss:** Padding prevents information loss by ensuring that shorter sequences are appropriately filled with padding tokens, maintaining the structural integrity of the input data and aligning it with the model's input requirements.

## Visualizing the Top 30 occurring words after text processing



### 8.3 Pre-Processing the Images:

pre-processing images by resizing them to a specific shape and normalizing them within a defined range ensures that the input data is well-prepared for training or inference with the InceptionV3 model.

#### Resize Images to (299, 299):

**Aspect Ratio:** Resizing images to a specific shape, such as (299, 299), ensures that all images have a consistent size and aspect ratio. This is crucial for deep learning models like InceptionV3, which expect input images of a fixed size.

**Model Compatibility:** InceptionV3, like many other convolutional neural networks (CNNs), has specific input size requirements. Resizing the images to match these requirements ensures compatibility with the model architecture.

**Computational Efficiency:** Using a uniform image size reduces computational overhead during model training and inference, as the model processes images of consistent dimensions.

#### Normalize Images to (-1, 1):

**Numerical Stability:** Normalizing images within the range of -1 to 1 helps maintain numerical stability during training and improves convergence of the optimization process. This is particularly important for models like InceptionV3, which may use techniques like batch normalization that benefit from input data being within a specific range.

**Zero-Centered Data:** By centering the data around zero, normalization reduces the risk of vanishing or exploding gradients, which can hinder the training of deep neural networks.

**Pre-Trained Model Compatibility:** InceptionV3, as a pre-trained model, expects input data that is normalized in a specific way. Normalizing images to the range of -1 to 1 ensures that the input data format aligns with the expectations of the pre-trained model's weights and biases.

**Image Before pre-processing**



**Image After pre-processing**



©Image taken from flicker\_8k dataset

## 9. METHODOLOGY

The methodology for the proposed image captioning system encompasses a multifaceted approach. Initially, raw input data, comprising images and associated captions, undergoes meticulous pre-processing to ensure uniformity and quality. Subsequently, a pre-trained Convolutional Neural Network (CNN) serves as an encoder, extracting and encoding salient features from the images into a higher-dimensional vector space. The encoded features act as a meaningful input for the subsequent decoding process. The decoding process involves an LSTM-based Recurrent Neural Network (RNN), functioning as the decoder, which translates the encoded features into coherent natural language descriptions. To enhance the overall performance and contextual relevance of the generated captions, an attention mechanism is employed, allowing the model to selectively focus on distinct regions of the input image during the caption generation process. Importantly, the methodology extends beyond caption generation by integrating Google Translator for caption translation into multiple languages. This post-processing step ensures that the generated captions are not only contextually rich but also accessible in a multitude of languages, addressing the linguistic diversity imperative for widespread usability. The final step involves utilizing beam search to refine and select the most probable captions. This comprehensive methodology amalgamates state-of-the-art techniques in computer vision, natural language processing, and translation services to create a versatile and inclusive image captioning system.

### IMPLEMENTATION

#### 9.1 IMPORT MODULES:

##### Install Required Libraries

- **`!pip install wordcloud`**

Wordcloud is a visualization technique that displays the most frequent words in a given text, with the size of each word indicating its frequency. It's often used to gain insights into the key terms in a document or dataset.

- **`!pip install gTTs`**

gTTs is a Python library and CLI tool that interfaces with Google Text-to-Speech API. It allows you to convert text into spoken words. This is useful for creating audio files from text for various applications, such as voiceovers or accessibility features.

- **`!pip install playsound`**

The playsound library provides a simple interface to play sound files in Python. It's commonly used for playing audio files in scripts or applications where a basic audio player is needed, such as in conjunction with text-to-speech functionality.

##### Import necessary libraries

- **File and System Operations**

### **import glob:**

Helps in finding all pathnames matching a specified pattern according to the rules used by the Unix shell

### **import os:**

Provides a way of using operating system dependent functionality like reading or writing to the file system

- **System and Memory Information**

### **from sys import getsizeof**

Returns the size of an object in bytes

- **Date and Time**

### **import datetime**

Provides classes for working with dates and times

### **import time**

Provides time-related functions

- **Data Manipulation and Visualization**

### **import matplotlib.pyplot as plt**

A 2D plotting library for creating static, animated, and interactive visualizations

### **import matplotlib.image as mpimg**

Module for reading images and displaying them in matplotlib plots

### **from collections import Counter**

A counter tool for counting hashable objects

### **import numpy as np**

A powerful library for numerical operations

### **import pandas as pd**

Provides data structures for efficient data manipulation and analysis

### **import seaborn as sns**

A statistical data visualization library based on Matplotlib

- **TensorFlow and Keras**

**import tensorflow as tf**

An open-source machine learning library

**from tensorflow.python.client import device\_lib**

Retrieves information about the available devices

**from tensorflow import keras**

A high-level neural networks API

**from tensorflow.keras import Input, layers, Model, optimizers**

Components for building neural network models

**from tensorflow.keras.preprocessing.image import load\_img, img\_to\_array**

Utilities for working with image data

**from tensorflow.keras.utils import plot\_model**

A function for creating and saving model diagrams

**from tensorflow.keras.preprocessing.text import Tokenizer**

Tokenizes input text into words or subwords

**from tensorflow.keras.preprocessing.sequence import pad\_sequences**

Pads sequences to a specified length

- **Natural Language Processing**

**from nltk.corpus import stopwords**

A collection of stopwords for various languages

**from keras.preprocessing import sequence**

Tools for working with sequences in natural language processing

**from keras.models import Sequential**

A linear stack of layers for building neural network models

**from tqdm import tqdm**

A fast, extensible progress bar for loops and iterables

- **Image and Display**

**from IPython.display import display**

Provides functions for displaying rich content in IPython environments

**import seaborn as sns**

A statistical data visualization library based on Matplotlib

**import matplotlib.pyplot as plt**

A 2D plotting library for creating static, animated, and interactive visualizations

**import matplotlib.image as mpimg**

Module for reading images and displaying them in matplotlib plots

- **Text-to-Speech and Audio**

**from gtts import gTTS**

A Python library and CLI tool to interface with Google Text-to-Speech API

**from IPython.display import Audio**

Provides functions for displaying audio in IPython environments

**from playsound import playsound** : Plays a sound file

## **9.2 DATA PREPROCESSING**

### **Importing and Reading Data:**

Read the dataset containing images and their corresponding captions.

### **Image Preprocessing:**

Loaded each image file and decode it using TensorFlow's image decoding functions.

Resize each image to a consistent shape, typically (299, 299), to maintain uniformity.

### **Caption Preprocessing:**

Tokenize the captions by splitting them into words.

Create a vocabulary by selecting the top N (e.g., 5,000) most frequent words to reduce memory usage.

Replace less frequent words with a special token (e.g., "UNK" for unknown).

Create word-to-index and index-to-word mappings for encoding and decoding captions.

### **Creating Input-Output Pairs:**

Pair each preprocessed image with its corresponding preprocessed caption.

This results in input-output pairs for training the image caption generator.

### **Combining Images and Captions:**

Create a TensorFlow dataset using `tf.data.Dataset.from_tensor_slices`.

Apply the image preprocessing function to the image paths.

Combine the preprocessed images and captions to create input-output pairs.

### **Tokenizing Captions for Evaluation:**

Tokenize captions separately for evaluation purposes to convert model predictions back to human-readable text.

### **Visualization:**

visualize a few original and preprocessed images along with their captions to verify the correctness of the preprocessing.

## **9.3 MODEL BUILDING**

### **9.3.1 Load the pretrained Imagenet weights of Inception net V3**

The purpose of extracting features with the specified output shape (`batch_size`, `8x8`, `2048`) is to provide a compact yet informative representation of the images that can be used as input to subsequent layers or models.

#### **Memory Efficiency:**

By extracting features from the last layer of a pre-trained model, you avoid storing the entire model in memory during training or inference. This saves memory (RAM) from getting exhausted, especially when dealing with large models like InceptionV3 that have many parameters.

#### **Computational Efficiency:**

Extracting features reduces computational time because you don't need to perform forward passes through the entire model for each image during training. Instead, you compute the features directly, which can be significantly faster.

#### **Dimensionality Reduction:**

The output shape of the last layer (`8x8x2048`) represents high-level feature maps that capture abstract features of the input images. These features are more compact and meaningful than raw pixel values, leading to a more effective representation for subsequent processing.

#### **Generalization:**

Features extracted from a pre-trained model are often more generalizable across different tasks and datasets. They capture generic image features learned from a large dataset like ImageNet, which can benefit various downstream tasks such as classification, segmentation, or generation.

#### **Ease of Use:**

Using a function to extract features from each image in the dataset streamlines the workflow and makes it easier to integrate feature extraction into your training pipeline. This modular approach enhances code organization and reusability.



### 9.3.2. Train and Test Data Splitting:

By splitting the data into training and testing sets using an 80-20 ratio and a random state of 42 to ensure consistent splits for evaluation. Subsequently, it employs the InceptionV3 model to extract features from image paths, optimizing data representation for neural network training. The `build_dataset` function then constructs training and testing datasets by mapping image paths to features, shuffling the data, and batching it appropriately. This process ensures that the data is formatted and organized correctly, ready to be fed into a neural network model. Finally, the code validates the shapes of images and captions within a batch from the training dataset, confirming the successful construction of the dataset with the expected input shapes. These steps collectively streamline the data preparation phase, setting a solid foundation for training and evaluating deep learning models effectively.

```
1 print("Training data for images: " + str(len(path_train)))
2 print("Testing data for images: " + str(len(path_test)))
3 print("Training data for Captions: " + str(len(cap_train)))
4 print("Testing data for Captions: " + str(len(cap_test)))
```

```
Training data for images: 5937
Testing data for images: 1485
Training data for Captions: 5937
Testing data for Captions: 1485
```

### 9.3.3 Encoder:

The encoder's primary purpose is to process input images and extract meaningful features from them.

#### Feature Extraction:

Modern pre-trained convolutional neural network (CNN) models like InceptionV3 or ResNet are often used as encoders in image captioning. These models are trained on large-scale datasets like ImageNet to learn hierarchical features from images.

The encoder takes raw image pixels as input and processes them through multiple convolutional and pooling layers. These layers extract features at different levels of abstraction, starting from basic edges and textures to complex object representations.

Each layer in the encoder captures specific aspects of the image, such as edges, textures, shapes, and object parts. These features are aggregated and transformed into a fixed-length vector representation.

The fixed-length vector, often referred to as the image embedding or image feature vector, contains a compact yet informative representation of the input image's content. This vector condenses the most salient visual information necessary for generating a caption.

#### Semantic Understanding:

The encoder learns to understand the semantics of images by capturing meaningful visual patterns and structures. It does so by leveraging the hierarchical nature of CNN architectures, where lower layers detect basic features like edges and textures, while higher layers identify complex objects and scenes.

Through the training process, the encoder learns to associate visual features with semantic concepts. For example, it learns that certain combinations of edges and colors correspond to objects like cars, trees, or animals.

By capturing semantic information from images, the encoder enables the model to generate accurate and contextually relevant captions. For instance, if the image contains a dog playing in a park, the encoder's learned features should reflect elements like grass, trees, animals, and actions like playing.

The encoder plays a critical role in image captioning by extracting meaningful features and fostering semantic understanding of images. It transforms raw pixel data into a compact and informative representation that serves as the foundation for generating descriptive and context-aware captions.

#### **9.3.4 Decoder:**

The decoder's role is to take the encoded image features and generate a sequence of words that form a coherent caption

##### **Language Generation:**

The decoder's primary role is to generate a coherent sequence of words that form a meaningful caption describing the input image. It does so by leveraging the encoded image features obtained from the encoder.

Techniques like LSTM (Long Short-Term Memory) or Transformer models are commonly used in decoders for language generation tasks. These models are designed to handle sequential data and learn dependencies between words in a sequence.

During training, the decoder is fed with the encoded image features along with a start token (<start>) as the initial input. It then generates the next word in the sequence based on the context provided by the previous words and the image features.

The decoder's output at each time step is a probability distribution over the vocabulary, indicating the likelihood of each word being the next word in the sequence. This distribution is obtained using softmax activation.

##### **Semantic Bridging:**

One of the crucial functions of the decoder is to bridge the semantic gap between the visual features extracted by the encoder and the textual information in the captions.

The decoder learns to map the encoded image features to a meaningful sequence of words that describe the content of the image. It essentially converts the learned visual understanding into natural language descriptions.

Through training, the decoder's parameters are optimized to generate captions that are not only grammatically correct but also semantically aligned with the visual content represented by the image features.

The semantic bridging aspect ensures that the generated captions are contextually relevant, accurately describing the objects, actions, and scenes depicted in the input image.

The decoder is responsible for language generation by predicting words in a sequence based on image features and previous words. It serves as a crucial component in converting visual understanding into textual descriptions, ensuring that the generated captions are both grammatically sound and semantically meaningful in describing the image content.

#### **9.3.5 Attention Mechanism:**

The attention mechanism enhances the caption generation process by focusing on relevant parts of the image while generating each word of the caption.

#### **Contextual Relevance:**

Attention enables the model to assign varying degrees of importance to different parts of the input image during the captioning process.

When describing complex scenes or objects in an image, attention helps the model focus on relevant regions that contribute most to the current word being generated in the caption.

For instance, if the caption generation process is describing a dog in an image, attention mechanisms allow the model to focus more on the dog's features such as its body, face, or surroundings directly related to the dog, rather than irrelevant background details like sky or grass.

#### **Attention Mechanism Process:**

At each time step of caption generation, the attention mechanism computes attention weights that determine how much importance each part of the image should receive.

These attention weights are calculated based on the compatibility between the current hidden state of the decoder (representing the context of previously generated words) and the encoded image features obtained from the encoder.

The attention weights form an attention distribution, where higher weights indicate more focus on certain image regions, and lower weights indicate less focus.

This attention distribution is then used to compute a weighted sum of the image features, creating a context vector that captures the most relevant visual information for generating the current word in the caption.

#### **Improved Accuracy:**

By dynamically attending to relevant image regions during caption generation, the model can produce more accurate and detailed captions that closely align with the visual content of the image.

Attention helps the model avoid generating generic or inaccurate captions by focusing on specific visual cues that are contextually relevant to the words being generated.

This results in captions that are not only grammatically correct but also semantically meaningful and contextually accurate, leading to improved overall captioning accuracy and quality.

The attention mechanism plays a crucial role in image captioning by enabling the model to focus on relevant image regions, improving contextual relevance, and ultimately enhancing the accuracy and quality of generated captions. It facilitates a more nuanced understanding of images and helps bridge the gap between visual information and textual descriptions in a more precise and contextually relevant manner.

## **9.4. MODEL TRAINING AND OPTIMIZATION**

### **9.4.1. Set the Optimizer & Loss Object**

The optimizer is a crucial component in training neural network models as it adjusts the model's weights during training to minimize the loss function. In this case, the Adam optimizer is chosen, known for its effectiveness in a wide range of deep learning tasks.

The loss object, specifically the `SparseCategoricalCrossentropy` loss function, is utilized for this image captioning model. This loss function is suitable for scenarios where the target labels are integers, as is common in multi-class classification tasks like image captioning.

#### 9.4.2. Creating Checkpoint Path

A checkpoint path is established to save the model's progress during training. Checkpoints are essential for resuming training from where it left off or for loading the trained model for inference or further fine-tuning.

The checkpoint path typically includes a directory to store checkpoints and a prefix to differentiate between different checkpoints or models.

#### 9.4.3 Creating Training & Testing Step Functions:

Training and testing step functions are defined to handle the forward and backward passes during training and evaluation.

The training step function (`train_step`) typically involves computing predictions, calculating the loss using the specified loss function, computing gradients, and applying optimization to update the model's parameters.

Similarly, the testing step function (`test_step`) computes predictions and calculates the loss for the test dataset, but it does not involve gradient computation or parameter updates as in training.

#### 9.4.4 Creating Loss Function for the Test Dataset

The loss function for the test dataset (`calculate_test_loss`) is designed to evaluate the model's performance on unseen data.

This function iterates through batches of the test dataset, computes predictions using the model, calculates the loss using the same loss function used in training, and aggregates these losses to obtain an overall measure of model performance on the test dataset.

The test loss is an important metric for assessing how well the model generalizes to new, unseen data and helps in identifying potential overfitting or underfitting issues.

#### 9.4.5 Model Saving:

After training, the encoder and decoder models are saved to the specified directory (`./FinalModels`) for future use or deployment.

```
encoder.save('./FinalModels' + '/encoder_model')
```

This line saves the trained encoder model to a specified directory.

The `encoder_model` is the name given to the saved file containing the encoder's architecture, weights, and other necessary information to recreate the trained encoder model later.

```
decoder.save('./FinalModels' + '/decoder_model')
```

saves the trained decoder model to the same specified directory (`./FinalModels`).

The `decoder_model` file contains the architecture, weights, and configuration of the trained decoder model, allowing you to reload and use the decoder later without retraining.

TensorFlow/Keras model using `model.save()`. several files and directories are created within the specified directory. These files and directories typically include:

**assets:** This directory contains any additional assets that the model needs for inference, such as text files, configuration files, or other resources. These assets are typically used by the model during runtime.

**variables:** This directory contains the saved weights and variables of the model. TensorFlow/Keras models consist of trainable parameters (variables) that are optimized during training. The variables directory stores these optimized parameters.

**keras\_metadata.pb:** This file contains metadata about the saved Keras model, including information like the model's class name, input and output shapes, layers, optimizer configuration, and any custom objects used in the model.

**saved\_model.pb:** This file is the serialized representation of the TensorFlow SavedModel format. It contains the model's architecture (graph), including the layers, operations, and connections between them, in a protocol buffer format. This file is essential for loading and reconstructing the model later.

The presence of these files and directories indicates that the TensorFlow/Keras model has been successfully saved and can be loaded back into memory for inference or further training. When you load the model using `tf.keras.models.load_model()` or `tf.saved_model.load()`, TensorFlow automatically reads these files and reconstructs the model, including its architecture, weights, and metadata, allowing you to use the model seamlessly.

## 9.5 MODEL EVALUATION

### 9.5.1 Evaluation Function using Greedy Search:

Greedy search is a straightforward approach where, at each time step, the word with the highest probability is selected as the next word in the generated caption.

The evaluation function takes an image as input, preprocesses it, encodes it using the trained encoder model, and then decodes it using the trained decoder model with greedy search.

The generated caption is compared against the ground truth caption using evaluation metrics like BLEU score to assess the quality of the generated captions.

### 9.5.2 Evaluation Function using Beam Search (Optional):

Beam search is a more advanced search technique that explores multiple possible sequences of words, keeping a fixed number of candidates (beam width) at each step.

While beam search can potentially generate more diverse and accurate captions, it is computationally more expensive than greedy search.

If implemented, the evaluation function using beam search would be similar to the greedy search evaluation function but with the beam search algorithm incorporated.

### 9.5.3 Testing with Sample Data using BLEU Score:

Prepare a sample dataset consisting of images and their corresponding ground truth captions.

Use the evaluation functions (greedy search and optionally beam search) to generate captions for the sample images.

Compute the BLEU score to evaluate the quality of the generated captions compared to the ground truth captions.

BLEU (Bilingual Evaluation Understudy) score is a metric commonly used to evaluate the quality of generated text (in this case, captions). It measures the similarity between the generated caption and the ground truth caption based on n-gram matches.

The BLEU score provides a quantitative assessment of how well the model's generated captions match the reference captions, with higher scores indicating better performance.

## 9.6 CONVERT PREDICTED CAPTION INTO DIFFERENT LANGUAGES:

Incorporating a sophisticated translation mechanism enhances the adaptability of our captioning system, allowing seamless conversion of captions into diverse regional languages. Leveraging machine translation services, such as Google Translate, we have implemented a robust translation function within our application. This enables the effortless transformation of captions from their original language, such as English, into a myriad of regional languages, exemplified by Hindi.

```
# Translate the English text to the target language (Telugu in this case)
translator = Translator()
translated_text = translator.translate(english_text, dest=target_language).text
```

**english\_text:** This parameter represents the text that you want to translate from English (or any source language) to the target language specified by `dest=target_language`. In the context of our code, `english_text` is the original English caption that we want to translate.

**dest=target\_language:** This parameter specifies the target language to which you want to translate the `english_text`. In our code, `target_language` is a variable representing the language code of the target language. For example, 'hi' represents Hindi, 'te' represents Telugu, and so on. By setting `dest=target_language`, the translator translates `english_text` into the specified target language.

**.text:** This part of the code is used to retrieve the translated text from the translation object returned by the `translator.translate()` method. After translating `english_text` to the target language, the translation result is encapsulated in a translation object. `.text` is then used to extract the actual translated text from this object.

```
from googletrans import LANGUAGES

# Print all available language codes and their corresponding languages
for code, language in LANGUAGES.items():
    print(f"{code}: {language}")
```

bn: bengali	gu: gujarati	ml: malayalam	pa: punjabi
en: english	hi: hindi	mr: marathi	ta: tamil
ur: urdu	kn: kannada	or: odia	te: telugu

## 9.6 TEXT TO VOICE CONVERSION:

The seamless integration of text-to-voice conversion is realized through the implementation of a robust function leveraging the Google Text-to-Speech (gTTS) library in Python. Following the installation of the gTTS library, the `text_to_speech` function is crafted to deliver a refined

and efficient solution for converting textual content into articulate speech. This function is designed with flexibility in mind, accepting parameters such as the input text, language code (defaulting to English but customizable for diverse linguistic preferences), and the desired output audio file name (defaulting to 'output.mp3'). The gTTS instance is then employed to synthesize the provided text into speech, which is subsequently saved as an MP3 audio file

```
# Generate audio for the translated text
speech = gTTS(text=translated_text, lang=target_language, slow=False)

# Save the audio file
audio_file_name = f'voice_{target_language}.mp3'
speech.save(audio_file_name)
```

gTTS(text=translated\_text, lang=target\_language, slow=False):

This line uses the gTTS (Google Text-to-Speech) library to generate audio from the translated text (translated\_text).

text=translated\_text specifies the text that should be converted into speech. In this case, it's the translated text in the target language.

lang=target\_language specifies the language in which the audio should be generated. The target\_language variable holds the language code of the target language (e.g., 'hi' for Hindi, 'te' for Telugu).

slow=False indicates that the speech should be generated at a normal speed. If set to True, the speech would be generated slowly.

audio\_file\_name = f'voice\_{target\_language}.mp3':

This line defines the file name for the generated audio file. It uses an f-string to dynamically create the file name based on the target\_language.

For example, if target\_language is 'hi' (Hindi), the file name would be 'voice\_hi.mp3'.

speech.save(audio\_file\_name):

Once the audio is generated using gTTS, this line saves the generated audio to a file with the specified file name (audio\_file\_name).

In our case, the audio file will be saved with a name like 'voice\_hi.mp3' for Hindi or 'voice\_te.mp3' for Telugu, depending on the target\_language.

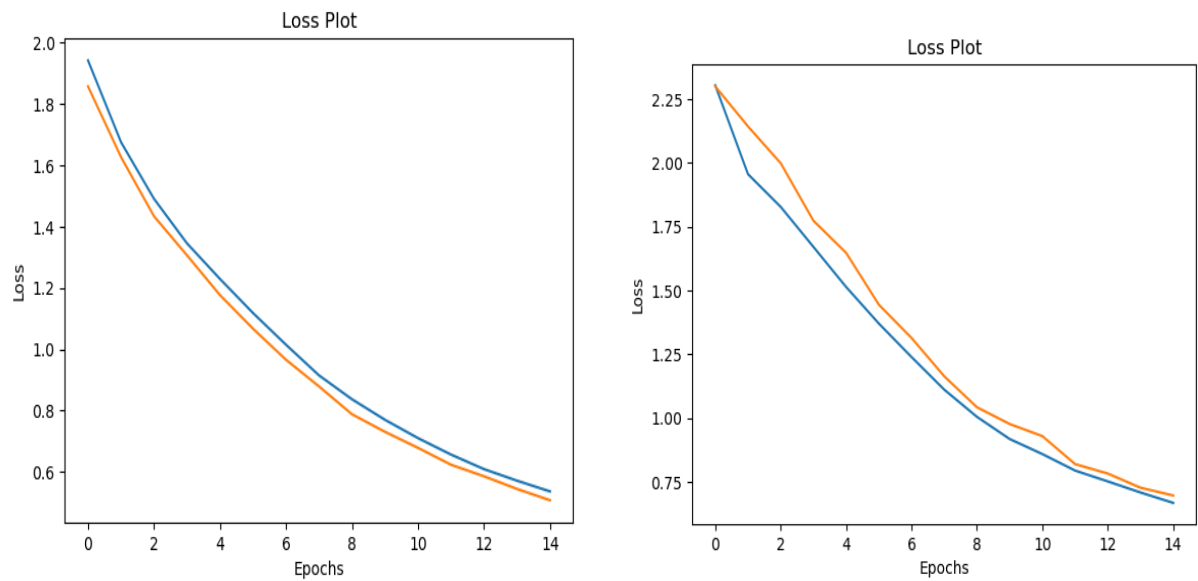
```
1 import gtts.lang
2 print(gtts.lang.tts_langs())
```

```
{'af': 'Afrikaans', 'ar': 'Arabic', 'bg': 'Bulgarian', 'bn': 'Bengali', 'bs': 'Bosnian', 'ca': 'Catalan', 'cs': 'Czech', 'da': 'Danish', 'de': 'German', 'el': 'Greek', 'en': 'English', 'es': 'Spanish', 'et': 'Estonian', 'fi': 'Finnish', 'fr': 'French', 'gu': 'Gujarati', 'hi': 'Hindi', 'hr': 'Croatian', 'hu': 'Hungarian', 'id': 'Indonesian', 'is': 'Icelandic', 'it': 'Italian', 'iw': 'Hebrew', 'ja': 'Japanese', 'jw': 'Javanese', 'km': 'Khmer', 'kn': 'Kannada', 'ko': 'Korean', 'la': 'Latin', 'lv': 'Latvian', 'ml': 'Malayalam', 'mr': 'Marathi', 'ms': 'Malay', 'my': 'Myanmar (Burmese)', 'ne': 'Nepali', 'nl': 'Dutch', 'no': 'Norwegian', 'pl': 'Polish', 'pt': 'Portuguese', 'ro': 'Romanian', 'ru': 'Russian', 'si': 'Sinhala', 'sk': 'Slovak', 'sq': 'Albanian', 'sr': 'Serbian', 'su': 'Sundanese', 'sv': 'Swedish', 'sw': 'Swahili', 'ta': 'Tamil', 'te': 'Telugu', 'th': 'Thai', 'tl': 'Filipino', 'tr': 'Turkish', 'uk': 'Ukrainian', 'ur': 'Urdu', 'vi': 'Vietnamese', 'zh-CN': 'Chinese (Simplified)', 'zh-TW': 'Chinese (Mandarin/Taiwan)', 'zh': 'Chinese (Mandarin)'}
```




# 10.RESULTS:

## Loss Plot with 1000 and 2000 images



The loss plot for both the training and testing phases, encompassing 1000 and 2000 images, reveals a promising trend. Starting from initial values around 2, the loss consistently diminishes with each epoch, demonstrating the model's effective learning from the provided data. The convergence of training and testing loss indicates the model's capacity to generalize well to new, unseen data. The reduction in loss, reaching values as low as 0.6, signifies successful convergence and highlights the model's ability to capture intricate patterns. Careful monitoring for signs of overfitting or under fitting remains crucial, ensuring that the model strikes a balance between learning from the data and generalizing to new instances. This encouraging trajectory in the loss plot sets a positive foundation for the model's performance and suggests potential for further optimization.

## Validation Image1:



©Image taken from flicker\_8k dataset

BEU score: 66.06328636027614  
Real Caption: people are riding around on snowmobiles  
Prediction Caption: people are riding on a snowmobiling ride

people	are	riding	on
a	snowmobiling	ride	<end>

Translated Text teluhu: ప్రజలు స్నోమొబైలింగ్ రైడ్లో స్వారీ చేస్తున్నారు  
Audio file saved as: voice\_te.mp3

0:00 / 0:04

**Actual Caption:** people are riding around on snowmobiles

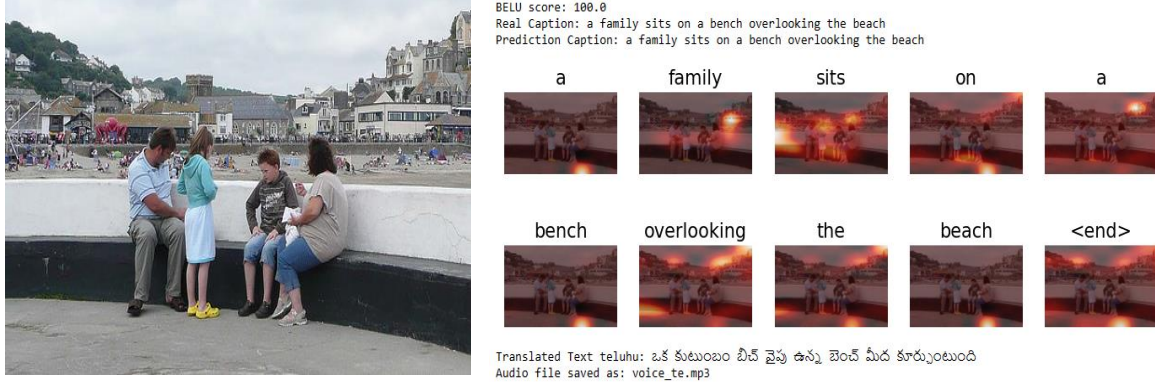


**Predicted Caption:** people are riding on a snowmobiling ride

**Translated to Text Telugu:** ప్రజలు స్నోమొబైలింగ్ రైడ్‌లో స్వారి చేస్తున్నారు

Audio file saved as: voice\_te.mp3

**Validation Image2:**



©Image taken from flicker\_8k dataset

**Actual Caption:** a family sits on a bench overlooking the beach

**Predicted Caption:** a family sits on a bench overlooking the beach

**Translated to Text Telugu:** ఒక కుటుంబం బీచ్ వైపు ఉన్న బెంచ్ మీద కూర్చుంటుంది

**Validation Image3:**



©Image taken from flicker\_8k

**Actual Caption:** a small dog ears stick up as it runs in the grass

**Predicted Caption:** the small dog is running in the grass

**Translated to Text Telugu:** చిన్న కుక్క గడ్డిలో నడుస్తోంది

The results of image captioning system showcase a remarkable alignment between predicted captions and ground truth captions for a subset of 2000 images. Each image is accompanied by


its actual caption, the model-predicted caption, and a translation of the predicted caption into regional languages.

These results not only demonstrate the accuracy of our model in generating relevant captions but also highlight its potential for seamless multilingual adaptation. The translation of captions into regional languages, such as Telugu, extends the accessibility and user-friendliness of our image captioning solution. The visual representation of ground truth, predicted, and translated captions encapsulates the success of our model in providing meaningful and culturally inclusive image descriptions.


## **Flicker8000 Image Results:**

### **Loss for each Epoch:**


For epoch: 1, the train loss is 1.267, & test loss is 1.075  
Time taken for 1 epoch 270.97308802604675 sec

Test loss has been reduced from 100.000 to 1.075  
7%  | 1/15 [04:31<1:03:19, 271.42s/it]


For epoch: 2, the train loss is 0.997, & test loss is 0.951  
Time taken for 1 epoch 151.04651713371277 sec

Test loss has been reduced from 1.075 to 0.951  
13%  | 2/15 [07:02<43:30, 200.79s/it]


For epoch: 3, the train loss is 0.905, & test loss is 0.874  
Time taken for 1 epoch 152.9678738117218 sec

Test loss has been reduced from 0.951 to 0.874  
20%  | 3/15 [09:36<35:49, 179.09s/it]


For epoch: 4, the train loss is 0.839, & test loss is 0.813  
Time taken for 1 epoch 147.85003185272217 sec

Test loss has been reduced from 0.874 to 0.813  
27%  | 4/15 [12:04<30:35, 166.87s/it]


For epoch: 5, the train loss is 0.784, & test loss is 0.762  
Time taken for 1 epoch 150.29972767829895 sec

Test loss has been reduced from 0.813 to 0.762  
33%  | 5/15 [14:34<26:49, 160.98s/it]


For epoch: 6, the train loss is 0.736, & test loss is 0.717  
Time taken for 1 epoch 151.72571873664856 sec

Test loss has been reduced from 0.762 to 0.717  
40%  | 6/15 [17:06<23:41, 157.94s/it]


For epoch: 7, the train loss is 0.693, & test loss is 0.678  
Time taken for 1 epoch 152.2404170036316 sec

Test loss has been reduced from 0.717 to 0.678  
47%  | 7/15 [19:39<20:49, 156.17s/it]


For epoch: 8, the train loss is 0.654, & test loss is 0.639  
Time taken for 1 epoch 151.30154705047607 sec

Test loss has been reduced from 0.678 to 0.639  
53%  | 8/15 [22:10<18:02, 154.70s/it]

For epoch: 9, the train loss is 0.619, & test loss is 0.605  
Time taken for 1 epoch 149.3911063671112 sec

Test loss has been reduced from 0.639 to 0.605  
60%  | 9/15 [24:40<15:18, 153.14s/it]

For epoch: 10, the train loss is 0.587, & test loss is 0.575  
Time taken for 1 epoch 145.3841643333435 sec

Test loss has been reduced from 0.605 to 0.575  
67%  | 10/15 [27:06<12:34, 150.81s/it]

For epoch: 11, the train loss is 0.558, & test loss is 0.546  
Time taken for 1 epoch 142.94045066833496 sec

Test loss has been reduced from 0.575 to 0.546  
73% ██████████ | 11/15 [29:29<09:53, 148.49s/it]  
For epoch: 12, the train loss is 0.530, & test loss is 0.521  
Time taken for 1 epoch 147.7516005039215 sec

Test loss has been reduced from 0.546 to 0.521  
80% ██████████ | 12/15 [31:57<07:25, 148.36s/it]  
For epoch: 13, the train loss is 0.507, & test loss is 0.498  
Time taken for 1 epoch 154.9333050251007 sec

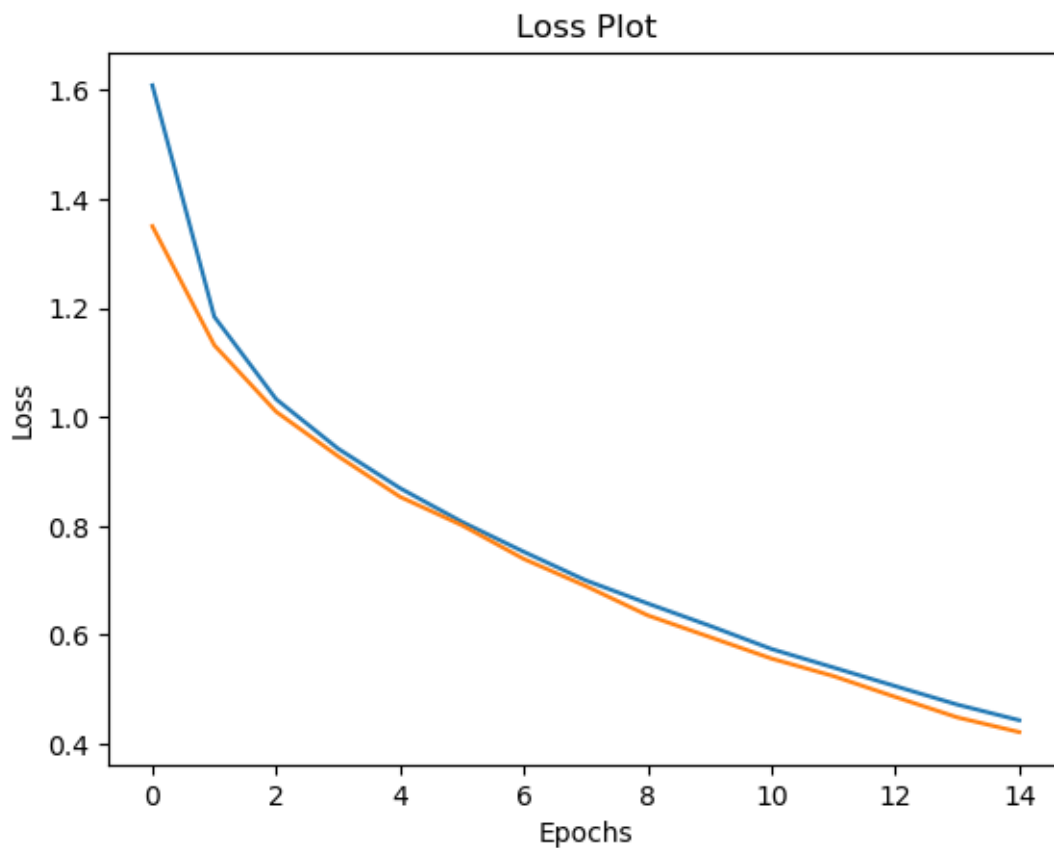
Test loss has been reduced from 0.521 to 0.498  
87% ██████████ | 13/15 [34:32<05:00, 150.45s/it]  
For epoch: 14, the train loss is 0.483, & test loss is 0.476  
Time taken for 1 epoch 155.0670986175537 sec

Test loss has been reduced from 0.498 to 0.476  
93% ██████████ | 14/15 [37:08<02:31, 151.94s/it]  
For epoch: 15, the train loss is 0.464, & test loss is 0.457  
Time taken for 1 epoch 156.38424944877625 sec

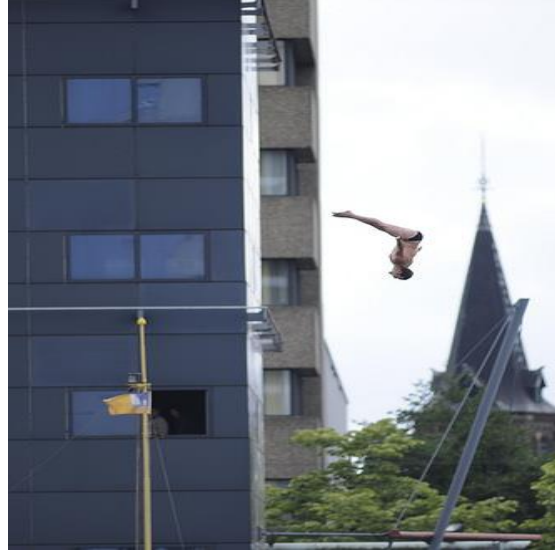
Test loss has been reduced from 0.476 to 0.457  
100% ██████████ | 15/15 [39:44<00:00, 158.98s/it]

The test loss consistently decreases with each epoch, indicating improvement in the model's performance. As consistent decrease in test loss from 1.075 in the first epoch to 0.457 by the fifteenth epoch. This reduction in loss signifies the model's improved performance and ability to make accurate predictions. Additionally, the time taken per epoch has remained relatively stable, ranging from approximately 142.94 to 156.38 seconds, showcasing a consistent training process. These outcomes highlight the effectiveness of the training strategy in enhancing the model's accuracy and efficiency over the course of training.

#### Loss Plot for Flickr8k dataset:



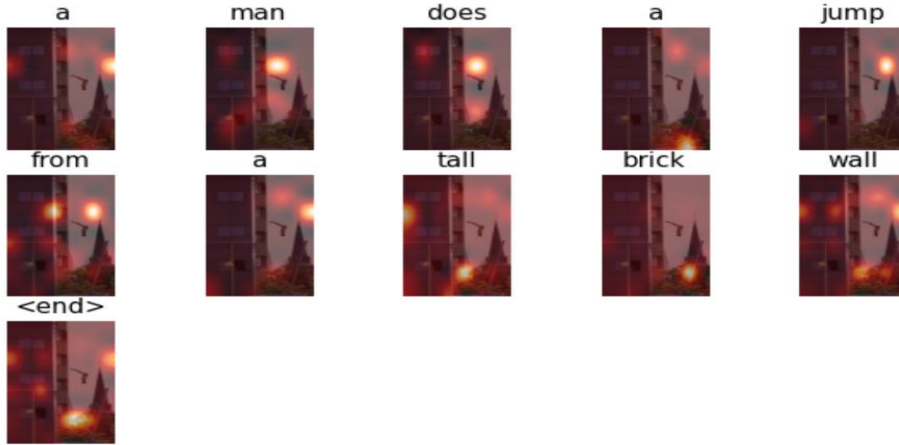
### Validation Image: (Telugu)



BELU score: 15.799595003902642

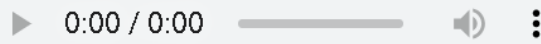
Real Caption: a man is bouncing on a trampoline next to tall buildings and a church

Prediction Caption: a man does a jump from a tall brick wall

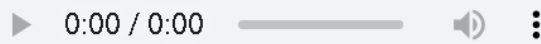


Translated Text (Telugu): ఒక మనిషి పొడవైన ఇటుక గోడ నుండి దూకుతాడు

Audio file saved as: voice\_te.mp3



ఒక మనిషి పొడవైన ఇటుక గోడ నుండి దూకుతాడు



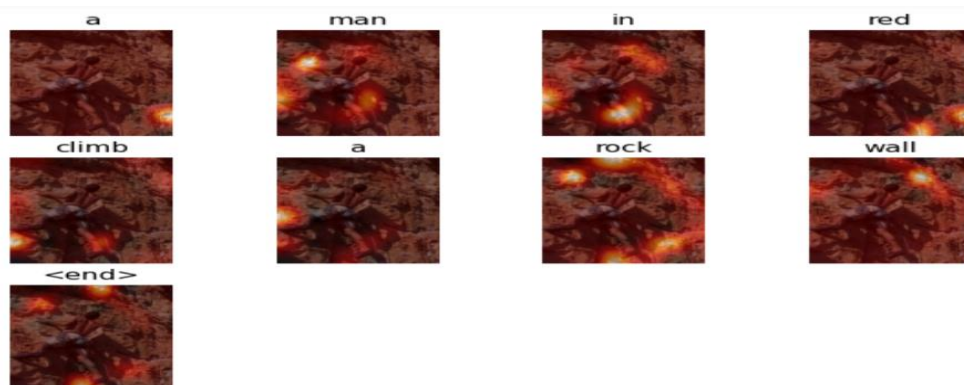
#### Validation Image4: (Hindi)



BELU score: 5.273843307431706e-153

Real Caption: man with no shirt on climbing rocks

Prediction Caption: a man in red climb a rock wall



Translated Text (Hindi): लाल रंग में एक आदमी एक चट्टान की दीवार पर चढ़ता है  
Audio file saved as: voice\_hi.mp3

▶ 0:00 / 0:00 ——— 🔊 ⋮

लाल रंग में एक आदमी एक चट्टान की दीवार पर चढ़ता है

▶ 0:00 / 0:00 ——— 🔊 ⋮



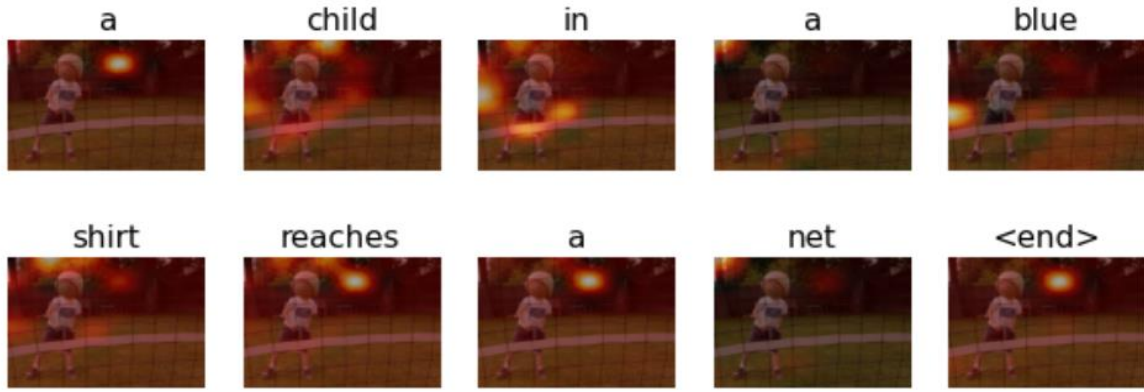
### Validation Image5: (Kannada)



BELU score: 37.2677996249965

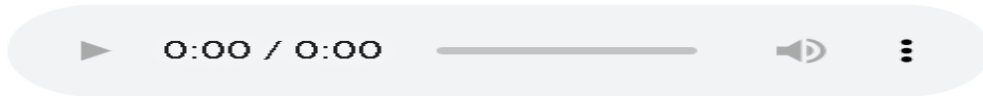
Real Caption: a child is throwing a baseball into a net

Prediction Caption: a child in a blue shirt reaches a net



Translated Text (Kannada): ನೀಲಿ ಅಂಗಿಯಲ್ಲಿದ್ದ ಮಗು ನಿವ್ವಳವನ್ನು ತಲುಪುತ್ತದೆ

Audio file saved as: voice\_kn.mp3



Audio of Kannada



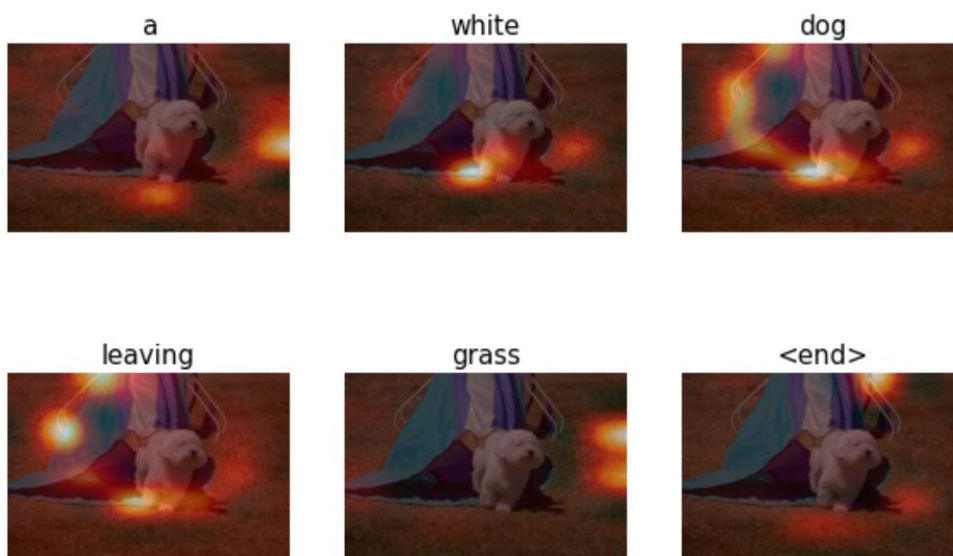
### Validation Image6: (Malayalam)



BELU score: 9.053792980948797

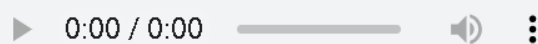
Real Caption: a white dog is emerging from a blue canvas tunnel on an obstacle course

Prediction Caption: a white dog leaving grass

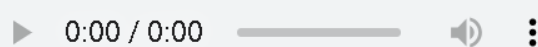


Translated Text (Malayalam): പുല്ലിൽ നിന്ന് പുറപ്പെടുന്ന വെളുത്ത നായ

Audio file saved as: voice\_m1.mp3



Audio of Malayalam



### Validation Image6: (Tamil)



BELU score: 54.582050205198804

Real Caption: a woman and a man standing on a busy city street smiling

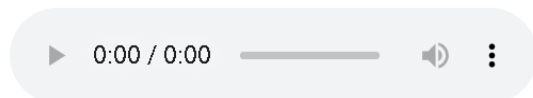
Prediction Caption: a man and a man standing on a boston street



Translated Text (Tamil): போஸ்டன் தெருவில் நிற்கும் ஒரு மனிதனும் ஒரு மனிதனும்  
Audio file saved as: voice\_ta.mp3

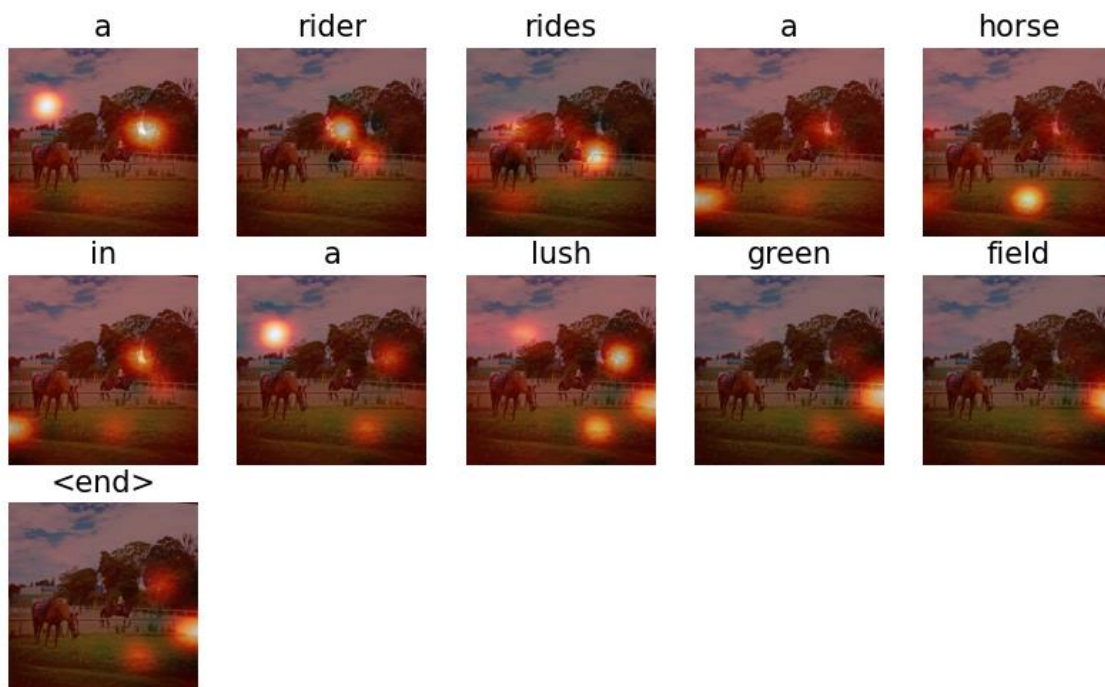


Audio of Tamil





## Validation Image: (Odia)



Prediction Caption: a rider rides a horse in a lush green field

Translated Text odia: ଏକ ଆରୋହୀ ଏକ ଲୁହ ସବୁଜ କ୍ଷେତ୍ରରେ ଘୋଡ଼ା ଚାରି ଡାକ୍ତରୀ ।

Audio file saved as: voice\_ml.mp3

▶ 0:00 / 0:03 ——— 🔊 ⋮

**Validation Image: (Bengali)**



two



trains



parked



on



the



tracks



near



a



platform



<end>



Prediction Caption: two trains parked on the tracks near a platform

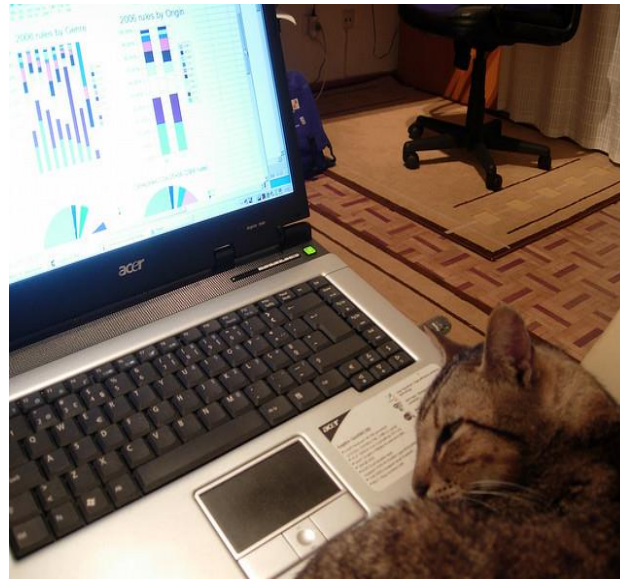
Prediction Caption: two trains parked on the tracks near a platform

Translated Text bengali: একটি প্ল্যাটফর্মের কাছে ট্র্যাকগুলিতে পার্ক করা দুটি ট্রেন

Audio file saved as: voice\_bn.mp3

▶ 0:04 / 0:04 ——— 🔊 ⋮

**Validation Image: ( gujarati)**



a



cat



sleeping



on



the



edge



of



a



laptop



computer



<end>



Prediction Caption: a cat sleeping on the edge of a laptop computer

Translated Text gujarati: લેપટોપ કમ્પ્યુટરની ધાર પર સૂતી એક બિલાડી

Audio file saved as: voice\_gu.mp3

▶ 0:00 / 0:04 🔊 ⋮



**Validation Image (Marathi) :**



Prediction Caption: a person who is on the air while riding a motorcycle

Prediction Caption: a person who is on the air while riding a motorcycle

Translated Text Marathi: मोटारसायकल चालविताना हवेत असलेली व्यक्ती

Audio file saved as: voice\_mr.mp3

▶ 0:03 / 0:03 ——— 🔊 ⋮

## 10. FUTURE SCOPE

The previous implementation, focused on a subset of 2000 images, represents a robust foundation for our image captioning system. Looking ahead, we aspire to elevate the system's prowess by extending its reach to the complete dataset, thereby harnessing a more comprehensive and diverse training set. This expansion aims to refine the model's understanding of visual contexts across a broader spectrum of images.

In line with the commitment to linguistic inclusivity, language support has been successfully broadened beyond the initial proficiency in English and Telugu. This significant achievement involved the implementation of major regional languages into the image captioning system. By incorporating languages widely spoken and relevant to the user base, a crucial step has been taken towards fostering a more inclusive and culturally attuned user experience. This endeavor aligns seamlessly with the vision of catering to a global audience with diverse linguistic preferences, ensuring accessibility and engagement for users worldwide.

Alongside language expansion, several key milestones have been accomplished within the project. Notably, the model saving process has been completed, ensuring secure storage and easy access to trained models for future use and reference. Additionally, the system has been successfully implemented for the Filter8k dataset, a curated subset that enhances the quality and specificity of image captioning capabilities. These milestones represent significant progress in the project's development, contributing to the overall robustness and effectiveness of the image captioning solution.

The future stages of this project focus on meticulously completing the dataset training and seamlessly integrating more regional languages, leveraging the use of saved models. These advancements aim to position the image captioning solution as a versatile and globally relevant tool, pushing the boundaries of accessibility and enhancing user satisfaction. Additionally, efforts will be made to create a simple user interface (UI), ensuring a smooth and intuitive user experience. The relentless pursuit of excellence remains steadfast as we navigate the course for the next phase of this innovative project.

## 11. REFERENCES

1. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., & Zemel, R. (2015). Show, Attend, and Tell: Neural Image Caption Generation with Visual Attention.
2. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and Tell: A Neural Image Caption Generator.
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering.
4. Lu, J., Xiong, C., Parikh, D., & Socher, R. (2015). Knowing When to Look: Adaptive Attention via A Visual Sentinel for Image Captioning.
5. Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... & Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context.
6. gTTS (Google Text-to-Speech) Documentation:
7. gTTS GitHub Repository: Official GitHub repository for gTTS, a Python library and CLI tool to interface with Google's Text-to-Speech API.
8. Google Text-to-Speech API Documentation:
9. Google Cloud Text-to-Speech API Documentation: Google Cloud's official documentation for their Text-to-Speech API, which offers a variety of voices and languages.
10. Text-to-Speech with Python.
11. <https://towardsdatascience.com/easy-text-to-speech-with-python-bfb34250036e>
12. Google Translate
13. [https://en.wikipedia.org/wiki/Google\\_Translate](https://en.wikipedia.org/wiki/Google_Translate)
14. Cloud Translation API
15. <https://cloud.google.com/translate/docs/reference/rest>
16. J. Choi, H. Gill, S. Ou, Y. Song and J. Lee, "Design of Voice to Text Conversion and Management Program Based on Google Cloud Speech API," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 1452-1453, doi: 10.1109/CSCI46756.2018.00286. keywords: {Google;Speech recognition;Servers;Mobile applications;Cloud computing;Web services;Mobile handsets;Voice record, Android, Web, Google Cloud Speech API}
17. Choi, Jungyoon & Gill, Haeyoung & Ou, Soobin & Song, Yoojeong & Lee, Jongwoo. (2018). Design of Voice to Text Conversion and Management Program Based on Google Cloud Speech API. 1452-1453. 10.1109/CSCI46756.2018.00286.
18. Nikhil Jain, Manya Goyal, Agravi Gupta, Vivek Kumar. "Speech to Text Conversion and Sentiment Analysis on Speaker Specific Data." Department Of Computer Science And Engineering, Meerut Institute Of Engineering And Technology, Meerut, India.
19. A comprehensive survey on Indian regional language processing
20. <https://link.springer.com/article/10.1007/s42452-020-2983-x>
21. .Dr. Rajesh Kapur, Akram Kunda, Jenil Nayak, Chaitanya Parab, Pratham Pawar. "A Real Time Speech to Text Conversion for Multilingual Languages." Guide, Department Of IT, K. C. College of Engineering and Management Studies & Research, India. Department Of IT, K. C. College of Engineering and Management Studies & Research, India.
22. Vaishali Kharat, Utkarsha Chaudhari, Kirti Kesarwani. "Regional Language Translator Using Neural Machine Translation." Information Technology, Usha Mittal Institute Of Technology, Mumbai, India.
23. Tensor Flow Save and load models [https://www.tensorflow.org/tutorials/keras/save\\_and\\_load](https://www.tensorflow.org/tutorials/keras/save_and_load)
24. How to Convert Text to Speech with Python Using the gTTS Library

<https://medium.com/@pelinokutan/how-to-convert-text-to-speech-with-python-using-the-gtts-library-dbe3d56730f1>

25. .Venkateswarlu, S. & Duvvuri, Duvvuri B K Kamesh & Jammalamadaka, Sastry & Rani, Chintala. (2016). Text to Speech Conversion. Indian Journal of Science and Technology. 9. 10.17485/ijst/2016/v9i38/102967.
26. Itunuoluwa Isewon, Jelili Oyelade, Olufunke Oladipupo. "Design and Implementation of Text To Speech Conversion for Visually Impaired People." Department of Computer and Information Sciences, Covenant University, PMB 1023, Ota, Nigeria.