VISION TO SOUND: ENHANCING ACCESSIBILITY TO VISUALLY IMPAIRED

A Project report submitted in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY IN

ELECTRONICS AND COMMUNICATION ENGINEERING

Submitted By

ALLI GOPI: BU21EECE0100161
BHOOMIKA N: BU21EECE0100494
Under the Guidance of
Dr. ARVIND KUMAR, ASSISTANT PROFESSOR

(Duration: Date/Month/Year to Date/Month/Year)

Department of Electrical,Electronics and Communication Engineering
GITAM School of Technology
GITAM(DEEMED TO BE UNIVERSITY)
(Estd. u/s 3 of the UGC act 1956)
NH 207, Nagadenehalli, Doddaballapur taluk, Bengaluru-561203 Karnataka, INDIA.

DECLARATION

I/We declare that the project work contained in this report is original and it has been done by me under the guidance of my project guide.

Name:

Date: **Signature of the Student**

NAME:
Alli Gopi : BU21EECE0100161
Bhoomika N: BU21EECE0100494

Department of Electrical,Electronics and Communication Engineering [14 Bold]
GITAM School of Technology, Bengaluru-561203

CERTIFICATE

This is to certify that (Student Name) bearing (Regd. No.:) has satisfactorily completed Mini Project Entitled in partial fulfillment of the requirements as prescribed by University for VIIIth semester, Bachelor of Technology in "Electrical, Electronics and Communication Engineering" and submitted this report during the academic year 2024-2025.

[Signature of the Guide] **[Signature of HOD**

Table of contents

Chapter 1: Introduction

1.1 Overview of the problem statement:

The "Vision to Sound: Enhancing Accessibility for the Visually Impaired" project aims to improve the daily lives of visually impaired individuals by converting visual information into auditory descriptions. Leveraging advanced technologies, the project focuses on:

1. **Image Processing and Object Detection:** Using convolutional neural networks (CNNs) to analyze and identify objects, scenes, and text from camera-captured images.
2. **Natural Language Processing (NLP):** Generating coherent, contextually relevant descriptions from the analyzed image data to make the information understandable to users.
3. **Embedded Systems Integration:** Developing a compact, portable system with low power consumption for real-time processing, incorporating essential hardware components like microcontrollers, cameras, and audio devices.
4. **Audio Synthesis and Output:** Converting text descriptions into audible speech using text-to-speech (TTS) technology, which is delivered through headphones or speakers.

1.2 Objectives and goals:

Objectives:

1. **Develop a Functional Prototype:** Create a system that effectively converts visual information into auditory descriptions, incorporating advanced machine learning and text-to-speech technologies.
2. **Achieve High Accuracy in Object Detection:** Utilize convolutional neural networks (CNNs) to ensure precise image processing and object recognition.
3. **Generate Contextual Descriptions:** Implement natural language processing (NLP) to produce clear, relevant, and coherent auditory descriptions of detected objects and scenes.
4. **Design a Portable and Efficient System:** Build a compact, low-power embedded system that ensures real-time processing and is suitable for everyday use.
5. **Conduct User Testing and Optimize Performance:** Test the system with visually impaired users, gather feedback, and refine the system to enhance its effectiveness and usability.

Chapter 2 : Literature Review

1. **Title of the paper: Image caption generation using Visual Attention Prediction and Contextual Spatial Relation Extraction**

   Year: 2023

   Authors: Reshmi Sasibhooshan, Suresh Kumaraswamy & Santhoshkumar Sasidharan

   Key Findings:

   Approach: The paper introduces a new method for generating image captions that combines a Wavelet-based CNN and LSTM, enhanced by attention mechanisms to focus on key details and relationships within images.

   Technique: By using wavelet decomposition and attention networks, the model captures detailed features and spatial relationships between objects, leading to more accurate and descriptive captions.

   Results: The model performs better than previous methods, achieving higher accuracy and relevance in captions across popular datasets like Flickr8K and MSCOCO.

2. **Title of the Paper: Image Captioning - A deep learning approach using CNN and LSTM Network**

   Year: 2023

   Authors: Preeti Voditel,Aparna Gurjar,Aakansha Pandey, Akrati Jain, Nandita Sharma

   Key Findings: Goal: The study aims to create a system that can automatically describe images by combining visual recognition (CNN) with sentence generation (LSTM), using the Flickr8k dataset.

How It Works: The model identifies important features in an image through a CNN, then uses an LSTM to turn those features into a descriptive sentence, effectively linking vision and language.

Results: The system shows good performance, with around 60% accuracy in matching words in its captions to reference captions, demonstrating its effectiveness in describing images.

3. **Title of the Paper: Automatic image captioning combining natural language processing and deep neural networks**

   Year: 2023

   Authors: Antonio M. Rinaldi, Cristiano Russo, Cristian Tommasino

   Key Findings:

   Approach: Combining natural language processing with computer vision to create models capable of automatically generating descriptive captions for images.

   Technique: Using deep neural networks, particularly convolutional neural networks (CNNs) for feature extraction and recurrent neural networks (RNNs) with attention mechanisms for sentence generation.

   Results: Enhanced accuracy and detail in the generated captions, making the model effective for real-world applications in automatic image captioning.

4. **A Comprehensive Survey of Deep Learning for Image Captioning**

   Year: 2019

   Authors: MD. Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga

   Key Findings:

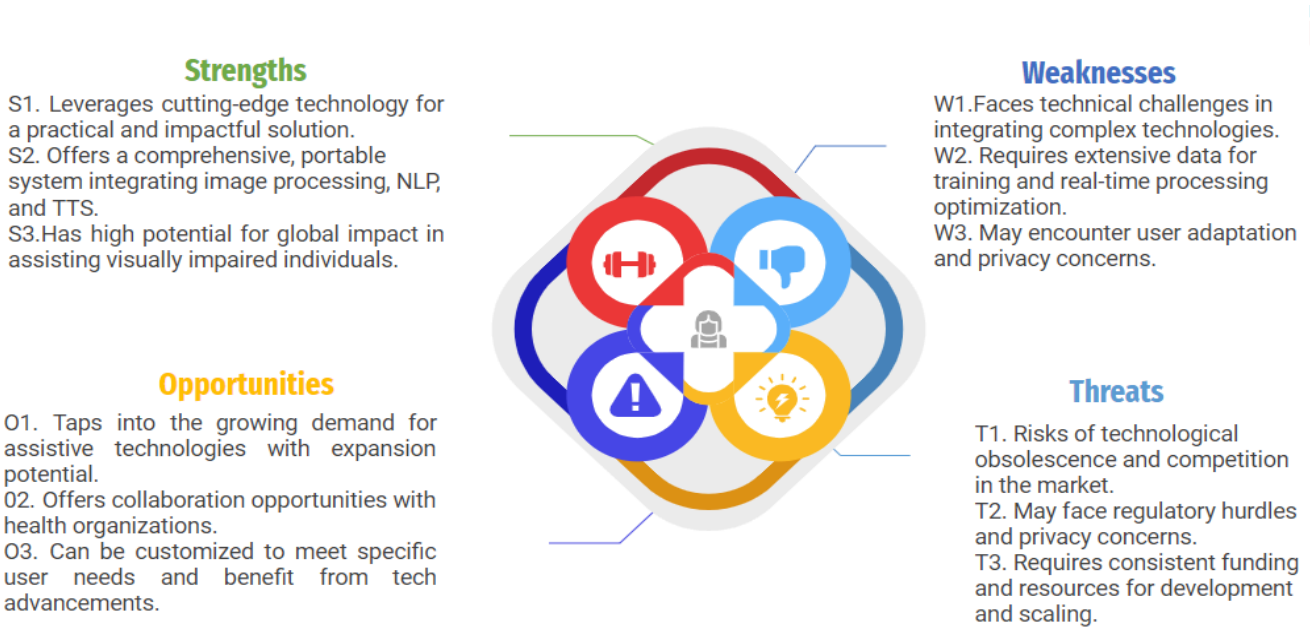
Deep Learning for Image Captioning: The paper reviews and categorizes deep-learning-based methods for generating image captions, focusing on novel caption generation techniques using CNNs, RNNs, and attention mechanisms.

Datasets and Metrics: It discusses widely used datasets like MS COCO and evaluation metrics such as BLEU and CIDEr for assessing caption quality.

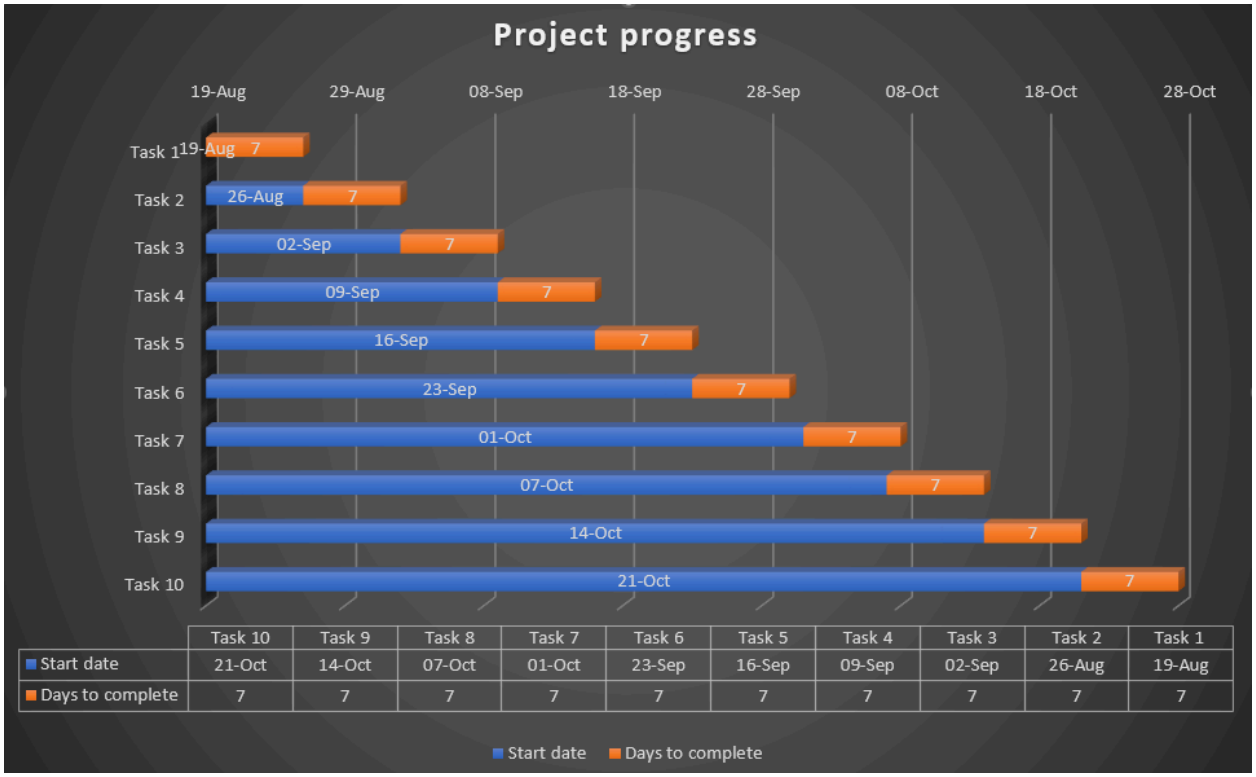
Comparison of Methods: The paper compares different deep learning techniques, analyzing their strengths, limitations, and performance in generating image captions.

Chapter 3 : Strategic Analysis and Problem Definition

3.1 SWOT Analysis



3.2 Project Plan - GANTT Chart



	Task 10	Task 9	Task 8	Task 7	Task 6	Task 5	Task 4	Task 3	Task 2	Task 1
Start date	21-Oct	14-Oct	07-Oct	01-Oct	23-Sep	16-Sep	09-Sep	02-Sep	26-Aug	19-Aug
Days to complete	7	7	7	7	7	7	7	7	7	7

Tasks	Start Date	Days to complete	Milestones
Task1	19-Aug	7	Abstract
Task2	26-Aug	7	Literature survey
Task3	02-Sep	7	Literature Survey
Task4	09-Sep	7	Yolo V4 - labelling the image
Task5	16-Sep	7	Working on Data set - Flickr8k
Task6	23-Sep	7	Working on CNN and LSTM
Task7	01-Oct	7	Combining CNN and LSTM
Task8	07-Oct	7	Deployment - image captioning
Task9	14-Oct	7	Testing
Task10	21-Oct	7	Documentation

3.3 Refinement of problem statement

Refined Problem Statement:
The project "Vision to Sound" aims to develop a comprehensive system that assists visually impaired individuals by translating visual scenes into descriptive audio cues. The refined objectives include:

Accurate Image Captioning: Developing a robust image captioning model that can not only identify objects within the image but also describe their relationships, actions, and context in a way that is meaningful to a visually impaired user.

Real-Time Processing: Ensuring that the system captures, processes, and converts images into audio descriptions in real-time or near-real-time, providing immediate feedback to the user.

Contextual and Relevant Descriptions: Refining the language and content of the captions to prioritize the most relevant details, avoiding information overload while ensuring that critical aspects of the scene are communicated effectively.

Natural and Intuitive Audio Output: Converting the generated captions into natural, clear, and easily understandable audio, using text-to-speech technology that accounts for tone, pace, and emphasis to convey the scene effectively.

Chapter 4 : Methodology

Data Collection and Preprocessing:

- **Dataset**: Utilized the Flickr8k dataset, consisting of 8,000 images and 5 corresponding captions per image.
- **Preprocessing**:
    - Resized and normalized images for uniform input to the model.
    - Tokenized and cleaned captions by removing unnecessary symbols and converting them into sequences of words.

Image Labeling using YOLO V4:

- **YOLO V4 Architecture**: Employed YOLO V4 for object detection due to its balance of speed and accuracy.

- **Labeling Process**: Labeled images using YOLO V4, identifying objects in images with bounding boxes and labels.
- **Result**: Generated labeled datasets that contain both the image and the objects detected, creating a foundation for further analysis.

Feature Extraction using CNN:

- **CNN Model**: Used a pre-trained Convolutional Neural Network (CNN) like VGG16 to extract features from each image.
- **Image Encoding**: Passed the images through the CNN to obtain feature vectors, which represent the key visual information in a lower-dimensional space.

Caption Generation using LSTM:

- **LSTM Architecture**: Designed a Long Short-Term Memory (LSTM) network to generate captions based on the extracted features.
- **Sequential Modeling**: The LSTM takes the image features from the CNN and sequentially generates text that describes the image, word by word.
- **Training**: The model was trained using image-caption pairs from the Flickr8k dataset to ensure accurate and context-aware captioning.

Integration of CNN and LSTM:

- Combined the CNN for feature extraction and the LSTM for sequential caption generation in an end-to-end model.
- The model was trained to minimize caption generation errors and to improve the contextual relevance of the captions.

4.1 Description of the approach

Data Collection and Preprocessing:

- **Dataset**: We utilized the Flickr8k dataset, a standard dataset for image captioning tasks, containing 8,000 images and 5 textual descriptions for each image. The diversity of images and captions allowed the model to learn from a wide range of scenarios.
- **Preprocessing**: Images were resized and normalized to ensure uniform input to the model. Caption text was tokenized (split into words) and cleaned by removing special characters and converting to lowercase, creating sequences that the model could process effectively.

Image Labeling using YOLO V4:

- **YOLO V4 Architecture**: YOLO V4 (You Only Look Once) is a fast and accurate object detection model. We used it to detect objects within the images by drawing bounding boxes around objects and assigning them labels. YOLO V4 was chosen for its real-time performance, which provides accurate results while maintaining high processing speeds.

- **Labeling Process**: Each image was processed using YOLO V4, resulting in labeled datasets where key objects were identified. This provided a strong foundation for integrating object detection with the caption generation system.

Feature Extraction using CNN:

- **CNN Model**: A pre-trained Convolutional Neural Network (CNN) such as VGG16 was used to extract key visual features from the images. CNNs are highly effective in recognizing patterns in images, such as shapes, textures, and objects, which are critical for understanding image content.
- **Image Encoding**: The images were passed through the CNN, which converted each image into a lower-dimensional feature vector, representing the essential information about the image. This feature vector was then passed to the caption generation component.

Caption Generation using LSTM:

- **LSTM Architecture**: We designed a Long Short-Term Memory (LSTM) network for generating textual descriptions of the images. LSTMs are a type of Recurrent Neural Network (RNN) capable of remembering long-term dependencies, making them ideal for sequential data like language.
- **Sequential Captioning**: Using the image features from the CNN, the LSTM network generated captions by predicting the next word in the sequence based on the current word and the image's context. This enabled the model to produce coherent and contextually relevant captions.
- **Training Process**: The CNN-LSTM model was trained on the image-caption pairs from the Flickr8k dataset. The model learned to associate specific visual features with corresponding words in the captions, gradually improving its ability to generate accurate and descriptive captions.

Integration of CNN and LSTM:

- The CNN and LSTM were integrated into a single end-to-end architecture. The CNN extracted image features, which were then fed into the LSTM for sequential caption generation. This architecture allowed for seamless conversion of image content into natural language descriptions, forming the core functionality of the system.

4.2 Tools and techniques utilized

YOLO V4 (You Only Look Once Version 4):

- **Purpose**: Real-time object detection and image labeling.
- **Technique**: Utilizes a deep convolutional network to detect and label objects within images, offering high speed and accuracy. YOLO V4 was chosen for its efficiency in detecting multiple objects in a single image frame.

Convolutional Neural Networks (CNN):

- **Purpose**: Feature extraction from images.
- **Technique**: A pre-trained CNN model (e.g., VGG16) was used to extract critical features from images, such as edges, shapes, and objects, which are later used for caption generation. CNNs are known for their ability to learn spatial hierarchies in visual data.

Long Short-Term Memory (LSTM) Networks:

- **Purpose**: Sequential caption generation.
- **Technique**: LSTM, a type of Recurrent Neural Network (RNN), was used to generate text captions from the image features extracted by the CNN. It models the sequence of words, capturing dependencies between them for meaningful, coherent captions.

Flickr8k Dataset:

- **Purpose**: Training the image captioning model.
- **Technique**: A dataset consisting of 8,000 images with corresponding textual descriptions (5 captions per image). This dataset was used to train the CNN-LSTM model to generate accurate captions by learning image-caption relationships.

4.3 Design considerations

Real-Time Performance:

- **Objective**: Ensure fast and accurate object detection and caption generation for real-time applications.
- **Consideration**: The choice of YOLO V4 for object detection allows for high-speed processing without compromising accuracy, making it suitable for real-time use. The CNN-LSTM model should also be optimized for low latency to avoid delays in caption generation.

Accuracy and Precision:

- **Objective**: Maintain high accuracy in both object detection and caption generation to provide meaningful and relevant output.
- **Consideration**: The models should be trained on diverse datasets (like Flickr8k) to generalize well to various environments and contexts. Fine-tuning the CNN and LSTM models is crucial for generating contextually accurate captions.

Scalability:

- **Objective**: Design the system to handle a growing number of images and increasing complexity over time.
- **Consideration**: The architecture should be scalable to handle larger datasets, more complex image scenarios, and additional features like multi-object captioning. The

model's design should support future integration with larger datasets (e.g., Flickr30k, COCO) for better performance.

Chapter 5 : Implementation

5.1 Description of how the project was executed

Dataset Preparation:

- **Flickr8k Dataset**: The project began by selecting the Flickr8k dataset, which contains 8,000 images with 5 corresponding captions each. This dataset was ideal for training the model for image captioning. The images were preprocessed by resizing and normalizing to ensure consistent input for the model.
- **Caption Preprocessing**: The captions were cleaned by removing special characters, converting text to lowercase, and tokenizing the words to convert them into sequences. This step helped in structuring the text data for the Long Short-Term Memory (LSTM) network.

Image Labeling using YOLO V4:

- **YOLO V4 Integration**: YOLO V4 was implemented for object detection in the images. YOLO V4's architecture was set up to detect multiple objects in images with high accuracy and speed.
- **Training**: YOLO V4 was either pre-trained on a large-scale dataset like COCO or fine-tuned using additional image-label pairs to improve detection accuracy. The model was able to identify objects in the images, drawing bounding boxes and labeling them, which laid the foundation for detailed image descriptions.
- **Output**: Labeled images with identified objects were generated, which were then passed on to the caption generation module.

Feature Extraction using CNN:

- **CNN Setup**: A pre-trained Convolutional Neural Network (CNN) such as VGG16 was used for feature extraction. The CNN processed the images and extracted essential visual features, such as shapes, textures, and patterns.
- **Image Encoding**: After extracting features, the CNN transformed the images into feature vectors, representing the key visual information in a lower-dimensional form. These vectors were then fed into the next stage for caption generation.

Caption Generation using LSTM:

- **LSTM Design**: A Long Short-Term Memory (LSTM) network was implemented to generate textual descriptions from the extracted image features. The LSTM was chosen for its ability to capture sequential dependencies, making it ideal for generating coherent sentences from image data.

- **Training the CNN-LSTM Model**: The CNN and LSTM were integrated into an end-to-end model. The CNN provided image features, which the LSTM used to generate captions one word at a time. The model was trained on the image-caption pairs from the Flickr8k dataset, allowing it to learn the relationships between visual features and the corresponding text.
- **Caption Generation Process**: During inference, for a given image, the CNN extracted features, and the LSTM generated the most probable words to form a description of the image.

5.2 Challenges faced and solutions implemented

Challenges faced:

Dataset Download and Management:

- **Challenge**: The large size of the Flickr8k dataset and associated captions made downloading and storing the data complex.
- **Impact**: Delays in accessing the dataset and storage constraints due to the volume of images.

Model Integration and Specific Image Handling:

- **Challenge**: Integrating YOLO V4 for object detection with CNN-LSTM for caption generation presented technical challenges.
- **Impact**: Difficulties in ensuring smooth data flow between models, particularly for complex images with overlapping objects or poor lighting.

Caption Generation Accuracy:

- **Challenge**: The CNN-LSTM model occasionally generated captions that were either too generic or failed to fully capture the context of the image.
- **Impact**: Reduced caption quality, especially in images with multiple objects or intricate scenes.

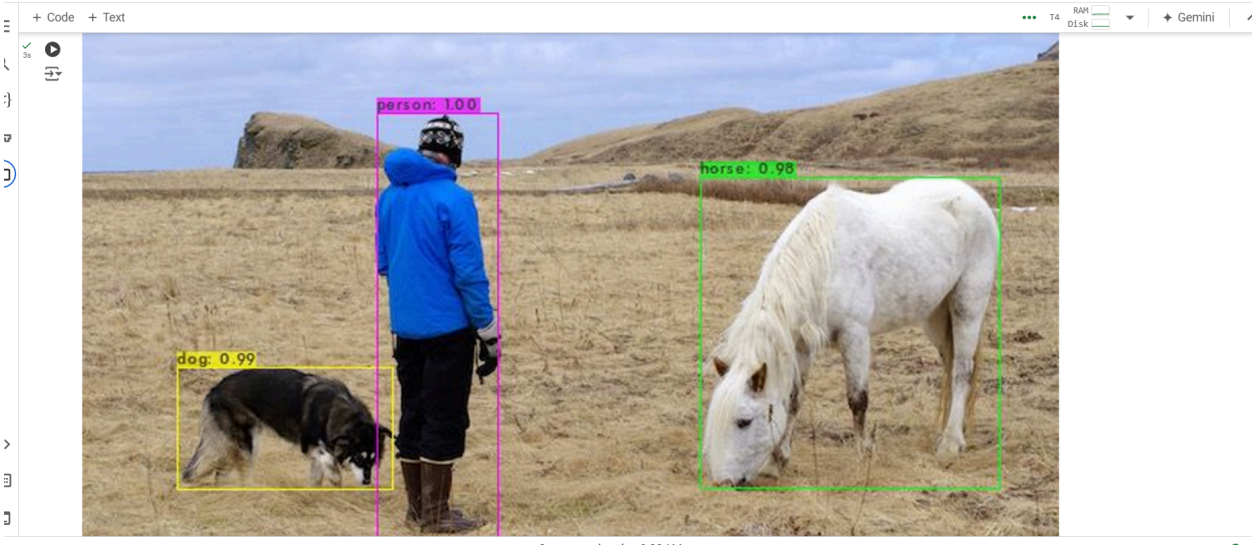
Solution Implemented:

- Utilized batch downloading methods and cloud storage options (Google Drive) to manage storage and ensure a smooth download process.
- Fine-tuned YOLO V4's detection parameters and applied data augmentation to enhance the model's robustness in diverse image conditions.

Chapter 6:Results

6.1 outcomes

1. Labeling of Image



2. Image Captioning

```
['startseq child in pink dress is climbing up set of stairs in an entry way endseq',
 'startseq girl going into wooden building endseq',
 'startseq little girl climbing into wooden playhouse endseq',
 'startseq little girl climbing the stairs to her playhouse endseq',
 'startseq little girl in pink dress going into wooden cabin endseq']
```

6.2 Interpretation of results

6.3 Comparison with existing literature or technologies

Chapter 7: Conclusion

Image Labeling Achieved: Successfully completed image labeling using YOLO V4, enabling accurate object detection in various images.
Advanced Caption Generation: Implemented image caption generation using a combination of Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, facilitating the transformation of visual data into meaningful text.

Utilization of Flickr8k Dataset: Leveraged the Flickr8k dataset to train the model, enhancing the quality and relevance of generated captions based on diverse image content.
Enhanced Accessibility: The integrated system aims to improve accessibility for visually impaired individuals by providing real-time audio descriptions of visual content, showcasing the practical application of advanced technologies in assistive solutions.

Chapter 8 : Future Work

Here write Suggestions for further research or development Potential improvements or extensions

References