Eye for Blind

Addressing Visual Accessibility Challenges: An AI Solution for Visually Impaired Users

Problem Statement

Visual impairment affects **285 million people worldwide**, with **39 million completely blind** according to WHO. These individuals face daily barriers in accessing visual content that dominates our digital world, where millions of images are shared across social media platforms.

Develop an AI-powered accessibility solution that converts any image into meaningful audio descriptions, enabling visually impaired individuals to understand visual content through detailed captions, bridge the accessibility gap in digital visual media.

The Digital Divide:

In today's image-centric social media landscape, visually impaired users cannot independently experience:

- Family and friends' shared moments
- Nature photography and scenic beauty
- Visual content that forms the backbone of social interaction
- Real-world images they encounter in their environment

Inspiration from Industry:

Facebook's pioneering accessibility feature demonstrated the potential of AI-powered image description, providing audio narration like *"This image may contain a dog standing with a man around the trees"* to help blind users understand posted content.

The project is an extended application of Show, Attend and Tell: Neural Image Caption Generation with Visual Attention paper.

Project Objective:

As per Evaluation Rubric, following objectives needs to be met for completion of project

1. **Data preparation**
2. **Modelling Building**
3. **Model Evaluation**
4. **Code readability and conciseness**
5. **Prediction**

Result: All steps executed successfully and available in Jupyter notebook.

Technical Approach:

Create an end-to-end pipeline using:

- **Computer Vision**: CNN-based feature extraction (InceptionV3)
- **Natural Language Processing**: RNN-based caption generation with attention
- **Audio Generation**: Text-to-speech conversion for accessibility
- **Deep Learning**: State-of-the-art encoder-decoder architecture

Core Innovation:

Transform visual information into **contextually relevant audio descriptions** by:

1. **Image Analysis**: Extract meaningful visual features
2. **Caption Generation**: Convert features to natural language descriptions
3. **Audio Output**: Provide immediate spoken descriptions

Impact:

This solution empowers visually impaired individuals to:

- Experience visual content independently
- Participate fully in image-based social interactions
- Navigate visual environments with AI assistance
- Enjoy equal access to visual media and information

The project addresses a critical accessibility need affecting hundreds of millions globally, using cutting-edge AI to create inclusive technology that bridges the gap between visual content and audio accessibility.

System Design:

- CNN-RNN Architecture with Attention

- Search Strategies (Greedy Search)

- Text-to-Speech Integration

- Comprehensive BLEU Score Evaluation

- Attention Visualization

This project aims to develop an AI system that describes images through speech, enabling visually impaired individuals to understand visual content.

Using the Flickr8k dataset, the model generates captions for images via a CNN and RNN architecture with an attention mechanism.

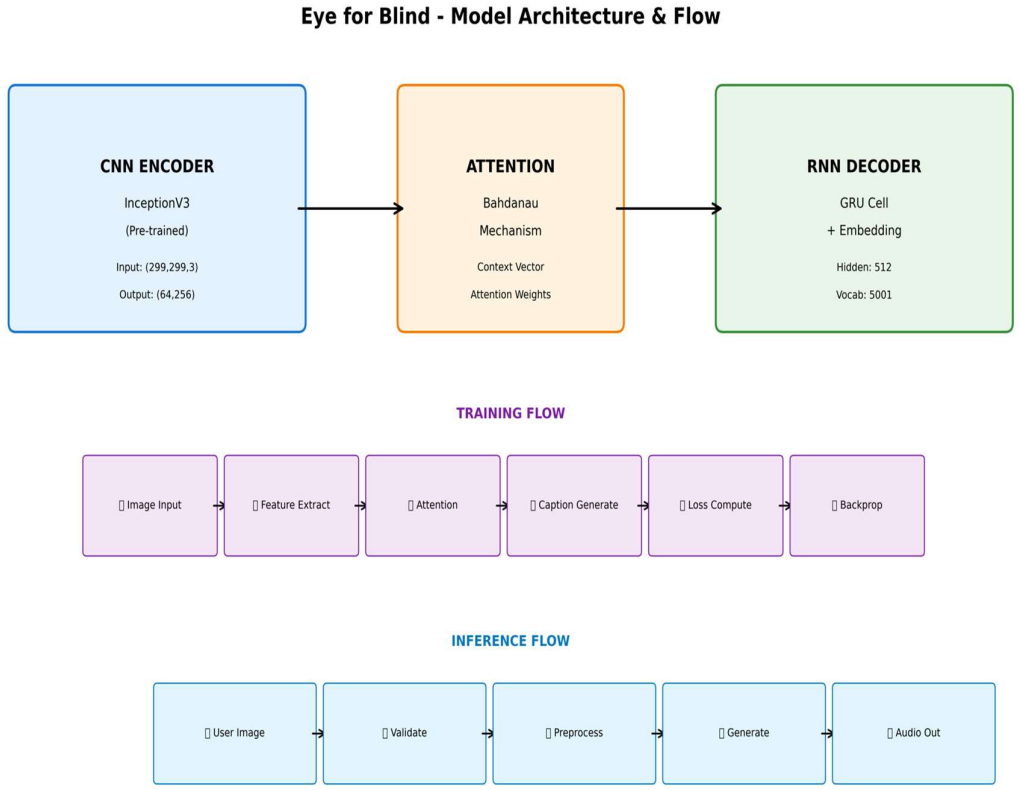
The CNN (Convolutional Neural Network) acts as an encoder, extracting key image features.

The RNN (Recurrent Neural Network) serves as a decoder, generating descriptive text (captions) word by word, guided by the attention layer to focus on relevant image regions.

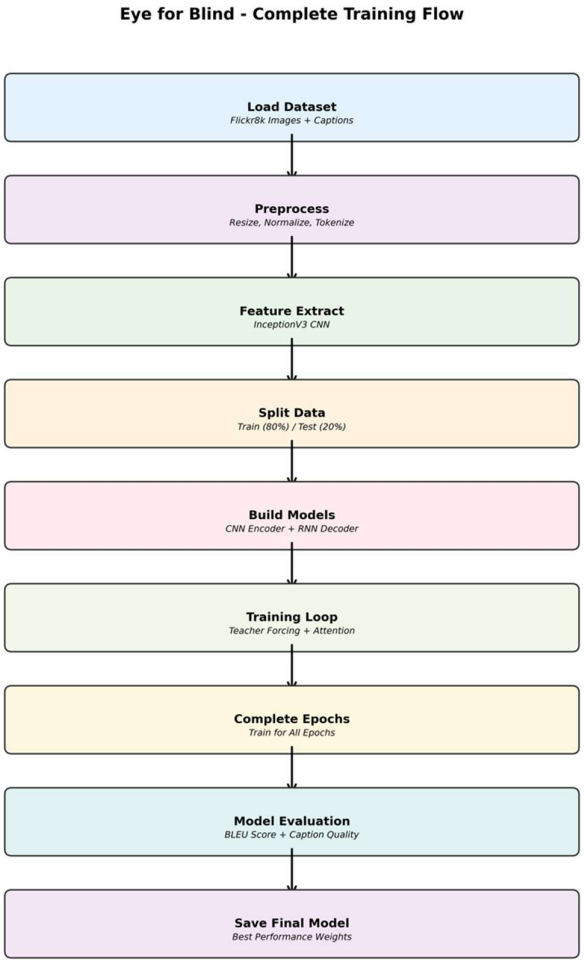
The generated text is then converted to audio output using a text-to-speech (TTS) library.

The project integrates Deep Learning (Computer Vision + NLP) to build an assistive system that translates images into spoken descriptions.

Architecture:



Training flow:



Implementation Summary

This project implements an image captioning model with attention and text-to-speech output. The key steps are:

1. **Data Loading and Preprocessing**:

   Data Understanding: Load the data and understand the representation.

   Data Preprocessing: Process both images and captions to the desired format.

   Steps:

   - Flickr8k dataset images and captions are loaded.
   - Captions are preprocessed: tokenized, filtered, and padded.
   - Images are resized and preprocessed for the InceptionV3 model.

- Image features are extracted using a pre-trained InceptionV3 model to save memory and computational time.
- Image paths and captions are split into training and testing sets.
- A data pipeline using tf.data.Dataset is created to efficiently load image features and captions in batches.

2. **Model Architecture**:

   Model Building: This is the stage where you will create your image captioning model by building Encoder, Attention and Decoder model.

   - An **Encoder** (CNN_Custom_Encoder) uses a fully connected layer to process the extracted image features.
   - An **Attention** mechanism (BahdanauAttention) calculates the attention weights, allowing the model to focus on relevant parts of the image during caption generation. Different attention mechanisms (Luong, Dot-Product) were also defined.
   - A **Decoder** (RNN_Custom_Decoder) uses a GRUCell and an Embedding layer to generate captions word by word, using the context vector from the attention mechanism and the previous word.

3. **Training and Optimization**:

   Train-Test Split: Combine both images and captions to create the train and test dataset and use optimizers and set parameters for training and test.

   - The model is trained using the Adam optimizer and Sparse Categorical Cross-Entropy loss with masking for padded sequences.
   - Teacher forcing is applied during training to guide the decoder with the true next word.
   - Training performed for defined number of epochs as it provided better results(predicted captions).
   - Early stopping is provided as choice for implementation to monitor test loss and prevent overfitting and saving the best model checkpoint but the Predicted captions are more near to ground truth without early stopping so it exists as choice but not used.
   - Training progress and loss curves are visualized.

4. **Model Evaluation and Prediction**:

   Model Evaluation: Evaluate the models using greedy search and BLEU score.

   - A greedy search function is defined to generate captions for new images.
   - The BLEU score is used to evaluate the quality of the generated captions against the real captions, with different weightings for n-grams.
   - A function is created to convert the generated captions into audio using gTTS.

- Attention maps are visualized to show which parts of the image the model is focusing on for each word in the generated caption.

Challenges faced

1. Compute power for model training.
2. Train and test loss optimization.
3. Model training and handling shape mismatches, symbolic graph errors, or gradient flow issues.
4. Choice between Training Completion(all Epochs) vs. Early Stopping: The training data output suggests overfitting (decreasing training loss vs. increasing test loss), this behaviour is expected in teacher forcing-based models where training uses ground truth inputs while testing relies on the model's own predictions.

    4.1 The decision to train for all epochs rather than implementing early stopping was based on comprehensive evaluation:
    4.2 Caption Quality Assessment: Generated captions demonstrated significantly higher semantic similarity to ground truth when trained for complete epochs
    4.3 Human Evaluation: Manual review of predicted vs. actual captions revealed that extended training produced more contextually accurate and meaningful descriptions
    4.4 Semantic Coherence: While predicted captions were not identical to ground truth, they captured essential visual elements more effectively with complete training
    4.5 Key Finding: Extended training yielded superior caption quality despite apparent loss divergence, confirming that semantic accuracy improved beyond what loss metrics alone indicated.

Future Improvements

1. GRU is used as RNN architecture and Bahdanau Attention as attention mechanism.

Model can be trained tested with LSTM RNN architecture and other attention(Dot-product attention , Luong or Multiplicative Attention) Mechanism and results can be compared for better results.

2. Beam Search is optional scope that can be covered.