Deep Learning Project Proposal Group-1

VisionVoice: Image Captioning App with Audio

Team Members-

1. Nihar Domala
2. Yash Manish Doshi
3. Nemi Makadia
4. Aravinda Vijayaram Kumar

Introduction

Image captioning, the task of automatically generating textual descriptions for images, has gained significant traction in recent years. By harnessing deep learning techniques, models can identify objects, actions, and scenes within images to produce coherent captions that closely resemble human descriptions. Image captioning has notable applications in accessibility, content organization, and automated content generation. When combined with audio output, captioning systems can further enhance accessibility for visually impaired users by providing spoken descriptions of visual content.

Objective

The goal of this project is to develop a robust image captioning model capable of generating descriptive captions for a wide range of images. Additionally, the project will integrate a text-to-speech (TTS) system that converts the generated captions into spoken audio output. This dual-functionality system will create an accessible, interactive experience where users can capture or upload an image and immediately receive an audio description. The project will utilize a combination of Convolutional Neural Networks (CNNs) for image feature extraction, Recurrent Neural Networks (RNNs) or Transformers for text generation, and TTS for producing high-quality audio output of captions.

Dataset

The COCO-2017 dataset is used in this project, containing 1,18,287 images with five to seven captions per image, covering a wide range of scenarios and contexts. This diversity in captions will enable the model to generalize well across different types of images, providing accurate and relevant descriptions.

Link: https://www.kaggle.com/datasets/awsaf49/coco-2017-dataset

Methodology

- **Image Feature Extraction**: Build a custom CNN from scratch to extract image features. This CNN will be trained to output a compact feature vector representing each image.
- **Caption Generation**: Use an RNN (LSTM or GRU) to generate captions based on the extracted image features. Feed the image feature vector as the initial input to the RNN, and train the RNN to predict captions word-by-word.
- **Text-to-Speech Conversion**: Develop a simple TTS model or integrate a pre-existing TTS framework to convert generated captions into audio output, making the captions accessible in spoken form.
- **Deployment**: Deploy the app on Streamlit Cloud or another hosting service for easy access.

Expected Outcome

The project will result in a fully functional system that generates accurate, descriptive captions for images and produces spoken audio output for each caption. This solution will have immediate applications in accessibility, particularly for visually impaired users, and could be used in interactive digital experiences across various industries.