Image Captioning Capstone Project

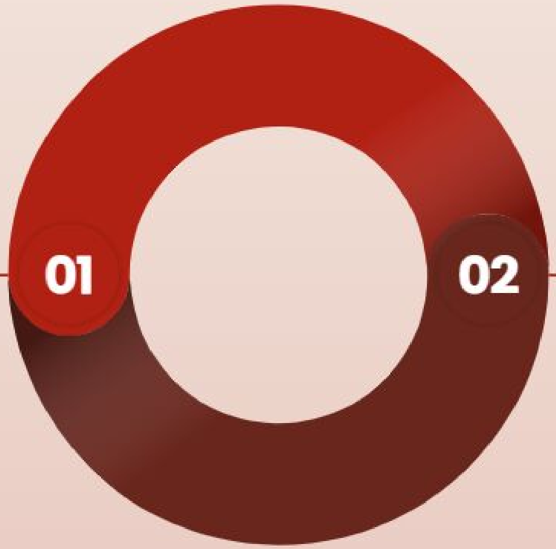
Done by **Rupesh**
Date October 2024

Table of Contents

- **Introduction**
  An overview of the project and its objectives.
- **Problem Statement**
  A clear definition of the problem being addressed.
- **Dataset Overview**
  Details about the dataset used for the project.
- **Project Architecture**
  The architecture and design of the project.
- **Model Training**
  The methodology and processes involved in training the model.
- **Challenges Faced**
  An outline of the main challenges encountered during the project.
- **Results**
  The outcomes and findings from the project.
- **Conclusion**
  A summary of the project and its implications.
- **Future Work**
  Potential directions for future research and development.

Problem Statement

Overview of Project Purpose and Constraints



Project Purpose

To develop an image captioning model using deep learning that generates descriptive captions for images.

Memory Constraints

Training was limited to 500 images due to memory constraints.

Dataset Overview

Insights into the Flickr8k Dataset

01

Flickr8k Dataset

Originally contains 8,000 images from Flickr, featuring five captions per image that focus on everyday scenes.

02

Dataset Source

The dataset is sourced from Kaggle, where 500 random images were selected to address memory limitations.

03

Caption Text Format

Each image has five different captions, formatted with and tokens.

04

Pre-processing Steps

The following steps are undertaken for pre-processing the dataset:

05

Image Extraction

Extracting images from the zip file.

06

Image Resizing

Resizing images to 128x128 pixels and normalizing pixel values.

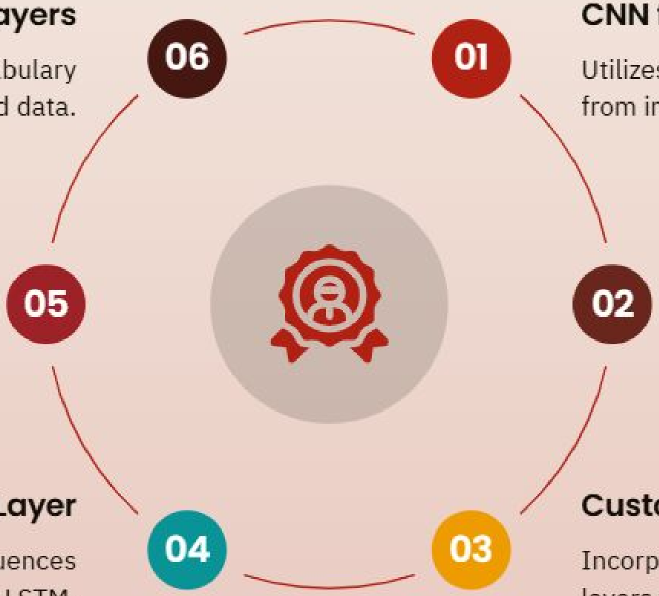
07

Caption Loading

Loading captions and filtering them for the selected 500 images.

Project Architecture

Overview of the Image Captioning Model



01 **CNN for Feature Extraction**

Utilizes Convolutional Neural Network to extract features from images.

02 **LSTM for Sequence Handling**

Employs Long Short-Term Memory networks to manage and generate caption sequences.

03 **Custom CNN Structure**

Incorporates Conv2D, MaxPooling, Flatten, and Dense layers for effective feature extraction.

04 **Embedding Layer**

Utilizes an embedding layer to process caption sequences before feeding them into the LSTM.

05 **LSTM Layer**

Processes the embedded sequences to generate meaningful captions.

06 **Dense and Output Layers**

Final layers are responsible for generating the vocabulary words based on the processed data.

Model Training

Overview of the Training Process and Setup

01 Input Image Preparation

Prepared input image sequences and padded caption sequences.

02 Caption Tokenization

Tokenized captions with a custom vocabulary, including , , and tokens.

03 Optimizer Used

Optimizer: Adam.

04 Loss Function

Loss function: Categorical Crossentropy.

05 Training Duration

Epochs: 30.

06 Batch Size

Batch size: 32.

07 Validation Data

Validation split: 20% of the data.

08 Training Constraints

Mention training constraints due to system resources and the impact on training duration.

Challenges Faced

Memory Constraints

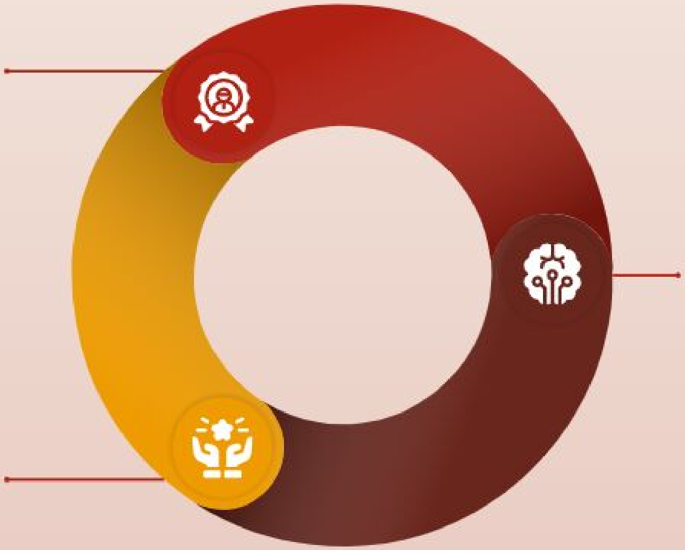
Limited to 500 images and smaller batch sizes to prevent system overload.

Model Accuracy vs. Resources

Managing trade-offs due to computational limitations.

Results and Model Evaluation

Evaluation Process Overview



Model Predictions vs True Captions

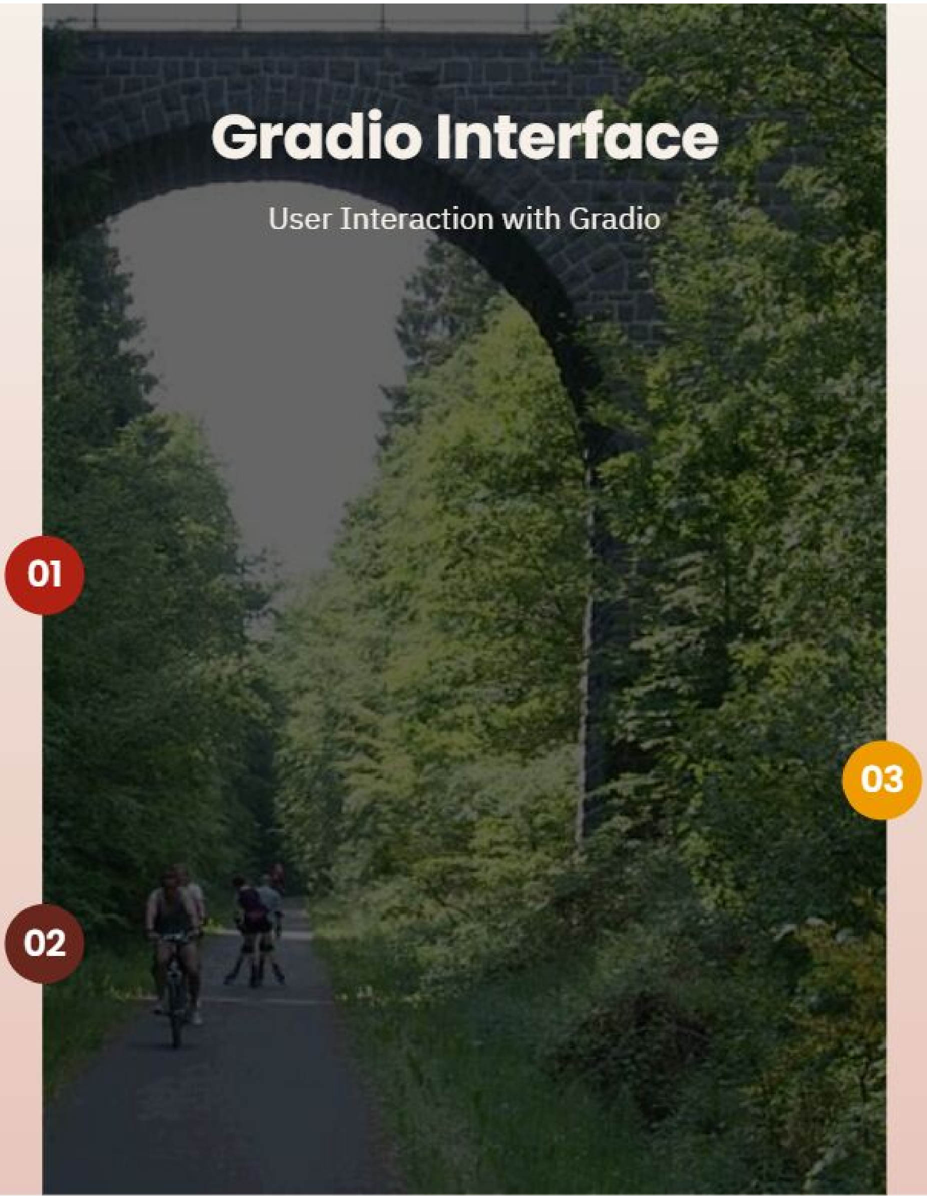
Generated captions from model predictions are compared to the true captions to evaluate performance.

Sample Image Display

Sample images are displayed with actual and predicted captions side-by-side for visual comparison.

Model Performance Summary

A summary of model performance is provided, focusing on challenges like model accuracy given limited data and memory constraints.



Gradio Interface

User Interaction with Gradio

Overview of Gradio Interface

01

The Gradio interface is designed for user interaction, providing a streamlined experience.

Interface Setup Process

02

Users can input an image into the interface and receive a generated caption as output.

Interactive Model Testing

03

Specific settings were utilized to test the model interactively, enhancing user experience.

Conclusion

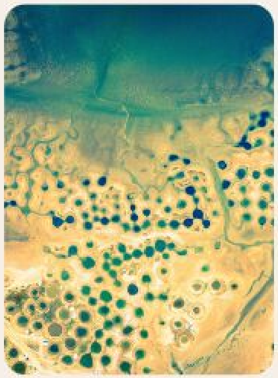
In this project, we developed an image captioning model that generates descriptive captions for images from the Flickr8k dataset. By leveraging a combination of Convolutional Neural Networks (CNNs) for feature extraction and Long Short-Term Memory (LSTM) networks for sequence generation, our model successfully learns to interpret visual content and articulate it in natural language.

Key accomplishments include:

Data Processing: Efficiently preprocessed and filtered a substantial dataset, ensuring high-quality input for model training.
Model Architecture: Implemented a custom CNN to extract image features and an LSTM to generate coherent captions, integrating these components effectively.
Performance Evaluation: Trained the model to produce relevant captions, demonstrating the potential of combining computer vision and natural language processing techniques.
The results indicate a promising capability for automatic image description, with applications in accessibility, social media, and content management systems. Future work may involve refining the model with larger datasets, exploring advanced architectures, and enhancing the quality of generated captions.

Future Work

Possible Improvements



Use of larger datasets or pretrained models

If resources allow, incorporating larger datasets or pretrained models can significantly enhance the model's performance and accuracy.

Model optimizations for memory efficiency

Implementing optimizations can lead to better memory usage, making the model more efficient and scalable.

Deployment for enhanced accessibility and performance

Strategic deployment can improve accessibility for users and boost the overall performance of the model.