IMAGE CAPTIONING MODEL

USING

DEEP LEARNING

TRAVEL GUIDE AI

Submitted by	Mentor
Sanchary Nandy	Sudheer Kumar Y

ABSTRACT

This project delves into the application of advanced machine learning techniques for the task of image captioning, specifically focusing on the rich diversity of Indian monuments. The study compares the effectiveness of two distinct architectures: a Transformer-based model and a CNN-RNN Hybrid model. Central to our approach is the use of the InceptionV3 encoder for extracting high-level features from monument images, coupled with custom decoders tailored to generate descriptive captions. Our investigation reveals compelling insights into the performance of these models. The CNN-RNN Hybrid model emerges as particularly adept at providing accurate captions for intricate and detailed monument images. This model leverages the sequential nature of Recurrent Neural Networks (RNNs) combined with the spatial awareness captured by Convolutional Neural Networks (CNNs), thereby excelling in capturing nuanced details and contextual relationships within images. In contrast, the Transformer-based model showcases its strengths in handling a wide array of monument images, demonstrating robust performance across diverse architectural styles and lighting conditions. The Transformer architecture's self-attention mechanism proves effective in capturing global dependencies within images, enabling it to generate coherent and contextually relevant captions across varied visual inputs.

Keywords: Machine learning, image captioning, Indian monuments, Transformer model, CNN-RNN Hybrid model, InceptionV3 encoder, custom decoders, comparative analysis, architectural diversity, cultural heritage, dataset evaluation

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 Requirements

This project focuses on building robust image captioning models. Key requirements include:

Comprehensive understanding of image captioning techniques.

Familiarity with the data science lifecycle: data acquisition, preprocessing, model development, evaluation, and deployment.

Proficiency in machine learning algorithms tailored for image data.

1.2 Goal and Objective

The primary goal is to develop and compare image captioning models for Indian monuments, providing handson experience with the data science lifecycle. The objective is to build models that can generate accurate and descriptive captions for monument images, enhancing understanding and appreciation of cultural heritage.

1.3 Data Science Process

- Data Preparation: Involves transforming raw image data into a clean and usable format, including handling missing values, removing outliers, and converting images into numerical representations.
- Data Exploration: Entails delving into the dataset to understand its characteristics and uncover hidden patterns, guiding feature engineering decisions.
- Data Modeling: Involves selecting or creating mathematical representations of the data relationships, training models, and making predictions on unseen data.
- Presentation and Automation:
- Focuses on transforming findings into easily digestible visuals and automating repetitive tasks within the data science lifecycle.

1.4 Introduction to Image Captioning

Image captioning involves generating descriptive textual content for images. This project leverages machine learning techniques to analyze and caption images of Indian monuments, providing insights into cultural heritage through automated descriptions.

CHAPTER 2: PROBLEM STATEMENT AND USE CASES

Problem Statement

Automated image captioning for Indian monuments faces several challenges:
Diverse architectural styles and features.
Variations in image quality and angles.
Need for contextually accurate and informative captions.

Use Cases

Enhancing educational materials with accurate descriptions of historical monuments.
Assisting visually impaired individuals by providing textual descriptions of images.
Supporting tourism by generating informative captions for monument images on travel websites.

CHAPTER 3: PROPOSED SOLUTION

3.1 Building Image Captioning Models

3.2 Data Overview

Dataset: Contains images of various Indian monuments with corresponding captions.

Images are divided into training, validation, and testing sets.

3.3 Image Preprocessing

Steps: Resize images, normalize pixel values, and convert images into feature vectors using the InceptionV3 encoder.

3.4 Text Representation

Need: Transform captions into numerical values for model training using techniques like tokenization and embedding.

3.5 Handling MultiClass Imbalance

Approach: Use techniques like oversampling/undersampling to balance the dataset and ensure accurate model training.

3.6 Train, Test, Validate Division

Process: Divide the dataset into training, validation, and test sets to evaluate model performance accurately.

3.7 ML Classification Algorithms

Models Used: Transformerbased model and CNNRNN Hybrid model.

CHAPTER 4: RESULTS AND DISCUSSION

4.1 Obtained Results

The image captioning models developed in this project, the Transformerbased model and the CNNRNN Hybrid model, have demonstrated promising results in generating accurate and informative captions for Indian monument images.

4.2 Interference

The performance of the two models highlights the tradeoffs between breadth and depth of understanding when it comes to image captioning. The Transformerbased model demonstrates a broader understanding of a wide range of monument images, while the CNNRNN Hybrid model excels in capturing the finer details and complexities of specific monument types.

CHAPTER 5: CONCLUSION AND FUTURE WORKS

5.1 Conclusion

This project has successfully developed and compared two stateoftheart image captioning models, the Transformerbased model and the CNNRNN Hybrid model, for the task of captioning Indian monument images. The results demonstrate the strengths and tradeoffs of each approach, providing valuable insights for future development and deployment of image captioning systems.

5.2 Future Works

To further enhance the performance and applicability of the image captioning models, the following future works are proposed:

- Explore additional data augmentation techniques to improve the models' ability to handle diverse and challenging monument images.
- Investigate the integration of external knowledge sources, such as architectural databases or historical information, to enrich the captions and provide more contextual understanding.
- Develop multimodal models that combine visual and textual inputs to generate more comprehensive and coherent captions.
- Conduct user studies to evaluate the practical usability and user experience of the generated captions, particularly for applications in education, tourism, and accessibility.
- Explore the deployment of the models in realworld scenarios and gather feedback to further refine and optimize the image captioning system.
- By addressing these future directions, the image captioning capabilities for Indian monuments can be further enhanced, contributing to the preservation and dissemination of cultural heritage through automated and informative descriptions.

REFERENCES

[1] https://arxiv.org/abs/1502.03044

[2] https://arxiv.org/abs/1411.4389

[3] https://arxiv.org/abs/1612.00563

[4] https://medium.com/@sirikrrishna99/automatic-image-captioning-using-streamlit-and-hugging-face-transformers-d3563edb5457

[5] https://www.analyticsvidhya.com/blog/2021/12/step-by-step-guide-to-build-image-caption-generator-using-deep-learning/